



OPEN ACCESS

EDITED BY

Maria Cutumisu,
McGill University, Canada

REVIEWED BY

Patrick Kyeremeh,
St. Joseph's College of Education, Ghana
Jennifer Dröse,
University of Paderborn, Germany

*CORRESPONDENCE

Eva Schultheis
✉ eva.schultheis@ph-freiburg.de

RECEIVED 28 November 2025

REVISED 13 February 2026

ACCEPTED 18 February 2026

PUBLISHED 08 April 2026

CITATION

Schultheis E, Leuders T, Reinhold F and
Loibl K (2026) Modeling and assessing
multiplicative operation
sense—validation of a test instrument
for 5th grade.
Front. Educ. 11:1756297.
doi: 10.3389/feduc.2026.1756297

COPYRIGHT

© 2026 Schultheis, Leuders, Reinhold
and Loibl. This is an open-access article
distributed under the terms of the
[Creative Commons Attribution License
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is
permitted, provided the original author(s)
and the copyright owner(s) are credited
and that the original publication in this
journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Modeling and assessing multiplicative operation sense—validation of a test instrument for 5th grade

Eva Schultheis^{1*}, Timo Leuders², Frank Reinhold² and
Katharina Loibl¹

¹Institute of Psychology, University of Education Freiburg, Freiburg, Germany, ²Institute of
Mathematics Education, University of Education Freiburg, Freiburg, Germany

Introduction: A considerable number of students enter secondary school with fundamental gaps in basic mathematical skills, particularly in their understanding of operations related to multiplication and division. The conceptual understanding of operations - operation sense - is the ability to relate situations (e.g., word problems) to mathematical-symbolic notations (e.g., calculations, equations) and vice versa and it is crucial for further learning in mathematics. In order to provide adaptive support for operation sense, the understanding or the lack thereof must be systematically assessed. Therefore, a sensitive assessment instrument specifically tailored to the operations of multiplication and division is needed, which can be used for a focused diagnosis and for evaluation of specific interventions in this domain. Thus, we developed an instrument for assessing multiplicative operation sense, which represents theoretically grounded levels of understanding operations in multiplicative situations.

Methods: The test was empirically validated in a pilot study ($N = 66$). Item responses were analyzed using general linear mixed models to investigate differences in solution rates across levels and to estimate the proportion of variance in item difficulty explained by the theoretically derived levels. Based on the pilot results, three items were revised and additional shortcomings of the study design were addressed in a main study ($N = 464$).

Results: In the pilot study, general linear mixed models showed that - as expected - the estimated solution rate decreases with increasing level and that 86% of the variance in item difficulty can be explained by the four theoretically derived levels of multiplicative operation sense. The main study showed 94% explained variance and significant mean differences between all levels.

Discussion: These findings support the validity of the instrument for assessing multiplicative operation sense and its usefulness for both research and practice.

KEYWORDS

competence model, formative assessment, multiplicative operation sense, test development, test validation

1 Introduction

Research shows that a considerable number of students in many countries still show substantial gaps in basic arithmetic knowledge and skills (Kasper et al., 2020; Mullis et al., 2020; Mullis et al., 2009; Sowder et al., 1998). In particular, the conceptual understanding of the basic arithmetic operations—or short: operation sense—is not sufficiently developed (Mullis et al., 2012), especially in the area of multiplication and division (Brown et al., 2010; Ehlert et al., 2013; Schulz et al., 2020). Operation sense can—in its widest sense—be defined as the ability to relate situations (e.g., word problems) to mathematical-symbolic notations (e.g., arithmetic calculations or equations) (Schulz et al., 2020). From a cognitive perspective adopted in this article, operation sense comprises the construction of situation models mentally representing the structure of the situation and the activation of mental models of mathematical operations that represent operations as reversible transformations on quantities.

In our region of Germany, elementary education ends with Grade 4, after which all students transit to secondary schools. Consequently, fifth-grade teachers encounter entirely new classes characterized by high student diversity, which can be challenging to assess and to address effectively (Gröhlich et al., 2009). Mathematics is a cumulative subject, and gaps in conceptual understanding can substantially hinder—or even prevent—further learning (Lamon, 2006; Wartha and Güse, 2009).

Understanding multiplication and division, for instance, plays a crucial role for further mathematical learning, for instance, when learning fractions, proportionality, percentages, or reasoning with functional relations (Baroody et al., 1999; Hackenberg and Tillema, 2009; Moss and London McNab, 2011; Slavit, 1998). Students who lack substantial understanding of multiplicative operations are, without a compensatory support, unlikely to achieve the requirements of mathematics instruction in secondary school (Hulbert et al., 2017; Schulz et al., 2017). Therefore, foundational skills, such as multiplicative operation sense, should be regularly assessed in Grade 5, with opportunities for reinforcement, consolidation, or, where necessary, a complete re-establishment of core concepts.

Because mathematics education is subject-specific, the diagnosis of missing knowledge in the understanding of multiplicative operations cannot simply be inferred from other areas, such as addition, which underscores the need for a dedicated diagnostic instrument targeting multiplicative operation sense at the transition to Grade 5.

Yet diagnostic instruments for operation sense are rather tailored to broadly screen proficiency among all arithmetic operations (Schulz et al., 2020) by written tests or they utilize individual diagnostic interviews (Moser Opitz et al., 2010) which cannot economically be employed for class assessments. Therefore, our aim is to develop a time-efficient sensitive test that specifically and reliably assesses different levels of students' multiplicative operations sense in a classroom setting. We consider such an instrument with its theoretically and empirically substantiated criteria as a prerequisite for evidence-based formative assessment and adaptive instruction.

For this purpose, a sensitive test specifically for multiplicative operation sense was developed and validated. Since a solid evidence-based validity argument is a prerequisite for the use of

such an instrument and for any intervention that may follow, the development and validation process of this instrument is described and evaluated in this article. Finally, possible applications in practice and research are discussed.

2 Theoretical and empirical foundation

As operation sense is understood as the ability to translate situations in real contexts (given as a picture, oral description, or text) into mathematical representations (i.e., numbers, arithmetic operations, expressions, equations, results) and vice versa (Schulz et al., 2020) it manifests itself in students' ability of solving word problems (Gravemeijer, 1997).

We deliberately choose the term “operation sense” (following Schulz et al., 2020) and not the terms “conceptual understanding” (Scheibling-Sève et al., 2020) or “multiplicative reasoning” (e.g., Jitendra et al., 2023) to emphasize the focus of our narrow conceptualization of the construct on the ability of solving multiplicative word problems.

2.1 Cognitive processes during the solution process of multiplicative word problems

Mathematical word problem solving requires the processing of linguistic information and numerical-mathematical operations. To capture this complexity, various cognitive models have been proposed that describe the processes involved in solving mathematical word problems. These models differ in how they conceptualize the interplay between linguistic and mathematical processes as well as the nature, role, and number of the mental representations assumed to underlie word problem solving (e.g., Reusser, 1989; Verschaffel et al., 2000; Borromeo Ferri, 2006; Verschaffel et al., 2020).

These models assume that solving word problems requires the construction or activation of two main internal representations which we focus on in our test development (see Figure 1): (1) the construction of a *situation model*, which specifies all objects, quantities and relations relevant for further processing, and (2) the activation of a *mental model* representing the appropriate operation and thereby facilitates the mathematization of the situation model.

First, a mental representation of the situation is built based on the textual basis: the so-called situation model. A viable situation model comprises an internal representation of all important elements and their relations. Constructing a situation model from the text provided in a word problem can be described by the processes typical for text comprehension (Kintsch and Greeno, 1985; Schnotz and Bannert, 2003): That is, one first has to understand the text sentence by sentence (local coherence building) (called textbase by Kintsch and Greeno, 1985), followed by the construction of an internal *representation of the situation*. This representation includes elements of the situation and their relations as described in the text. Some elements and relations are already given at the sentence level;

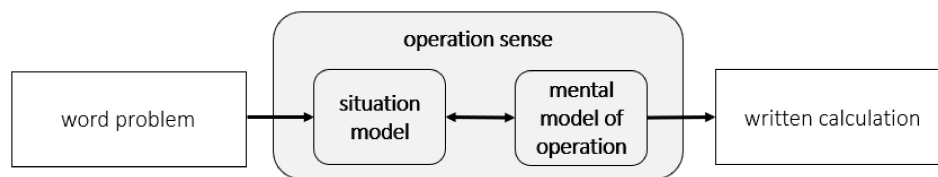


FIGURE 1

Process model highlighting the two main internal representations comprising operation sense in solving word problems.

others must be inferred from the semantics of the whole text (global coherence building). Such inferences draw on prior knowledge of the world and on schemata from long-term memory; schemata can be considered as generalized situation knowledge stored in long-term memory that are applied to the existing situation and generate further information (Anderson, 2018; McVee et al., 2005; Nesher and Hershkovitz, 1994; Piaget, 1976)

Second, the activation of an appropriate mental model of the operation is required. Having activated an appropriate mental model allows students to formulate a *mathematical procedure or description*, which can be enacted mentally or represented symbolically, e.g., by writing down the calculation as arithmetic expression for a division or a multiplication or by expressing this calculation verbally.

In mathematics education, mental models of operations are referred to by different terminologies and varying definitions (Fischbein et al., 1985; Jitendra et al., 2023; Marshall, 1995; Vom Hofe and Blum, 2016)—for the purpose of this study we follow the idea that basic arithmetic operations can model the numerical structure of situations (Greer, 1994; Verschaffel et al., 2000). Each arithmetic operation can be recontextualized in a variety of situations and can, vice versa, be used to mathematize a variety of situations (Schulz et al., 2020). Some of these different situations share the same generalized basic situation that requires a common operation. These generalized basic situations for an operation can be interpreted as a distinguishable mental model for the specific operation. A mental model can be understood as a coherent interpretation of a phenomenon or concept that typically encompasses a set of rules and constraints. The process of determining the operation required to solve a problem does not occur directly; instead, it is mediated by the model that is activated in response to the situation (Fischbein et al., 1985 as cited in Schulz et al., 2020, p. 429). In this article, we use the term mental model with this narrow meaning, similar to the concept of “Grundvorstellungen” (cf. Prediger, 2008; Schulz et al., 2020; Vom Hofe and Blum, 2016) in the German tradition.¹ The mental models relevant for our research, can be found in the literature (Greer, 1994, 1997; Schulz et al., 2020) as follows:

Mental models pertaining to the operation of multiplication are equal groups (static and dynamic), multiplicative (or proportional) comparisons, and Cartesian product (for more details and explanations on mental models see the table in Supplementary

Appendix). Mental models pertaining to the operation of division are those for partitive division (“sharing”) and for quotative division (“grouping”); partitive division corresponds to dividing by the multiplier, quotative division corresponds to dividing by the multiplicand.

The notion of conceptual field (Vergnaud, 1994) provides a framework for linking and coordinating these mental models: the more different typical situations a learner associates with appropriate mental models, the more stable their understanding of the underlying mathematical structure becomes. We consider multiplication and division not to be strictly distinct, but aspects of a common more complex integrated conceptual field for multiplicative situations (Vergnaud, 1994). Each multiplicative situation is determined by three values: the total quantity, the number of portions, and the size of a portion. Depending on what is required in each situation, a multiplication or a division must be calculated (Schulz and Wartha, 2021). Thus, multiplication and division as inverse operations are connected in the children’s multiplicative conceptual field. Accordingly, with understanding of multiplicative operations or multiplicative operation sense we refer to the application of mental models in both types of situations: multiplication and division.

When solving word problems, students often struggle because they are unable to identify the appropriate mental model of the operation required by the given situation, or use superficial and incorrect solution strategies, such as relying on keywords. Such superficial strategies indicate that students either did not build a situation model at all, or that their situation model is incorrect or restricted (Mayer and Hegarty, 1996; Verschaffel et al., 2000). Therefore, children’s responses to word problems are considered to reveal their understanding of operations, i.e., their level of operation sense.

2.2 Development of a test: difficulty generating dimensions of multiplicative word problems

To effectively promote operation sense and to be able to make reliable statements about the effect of any intervention, it is important and not at all trivial to have a sensitive and economic test instrument.

Existing instruments are either designed to be administered in an individual interview setting and have students explain their reasoning (e.g., The Diagnostic Interview from the New Zealand “Numeracy Project” by the Ministry of Education, 2008), which are not scalable to whole classrooms, or they test the construct of multiplicative operation sense as a subdomain of more

¹ This differs from an alternative usage in cognitive science, where mental models are considered internal *ad hoc* representations of a situation as, e.g., Johnson-Laird (1983). A mental model in our sense is closer to schemata, described by Marshall (1995) and Jitendra (2019).

general mathematical competencies such as BASIS-MATH 4–8 (Moser Opitz et al., 2010) or the Booker Screening Test (Booker, 2011). These well-established test instruments may be used as a screening or baseline measure, they offer the possibility of individually locating learners on an arithmetic-related performance scale, indications of dyscalculia and supplementary qualitative analyses. However, they do not provide sufficiently differentiated information on students' developmental levels and learning needs in multiplicative operation sense to inform subsequent interventions. Other instruments focus only on a specific subdomain of multiplicative operation understanding (Royar, 2013) or go new ways by seeking to assess multiplicative reasoning by using visual models as items without using canonical representation of multiplication and division (Kosko, 2019). Considering these specific strengths and boundaries of existing instruments, we see the need for an instrument that can be administered in whole classrooms and that provides precise information on students' skills that can guide subsequent instructional support. We seek an instrument which allows testing different levels of multiplicative operation sense to be relevant and easy applicable in the diagnostic practice at school.

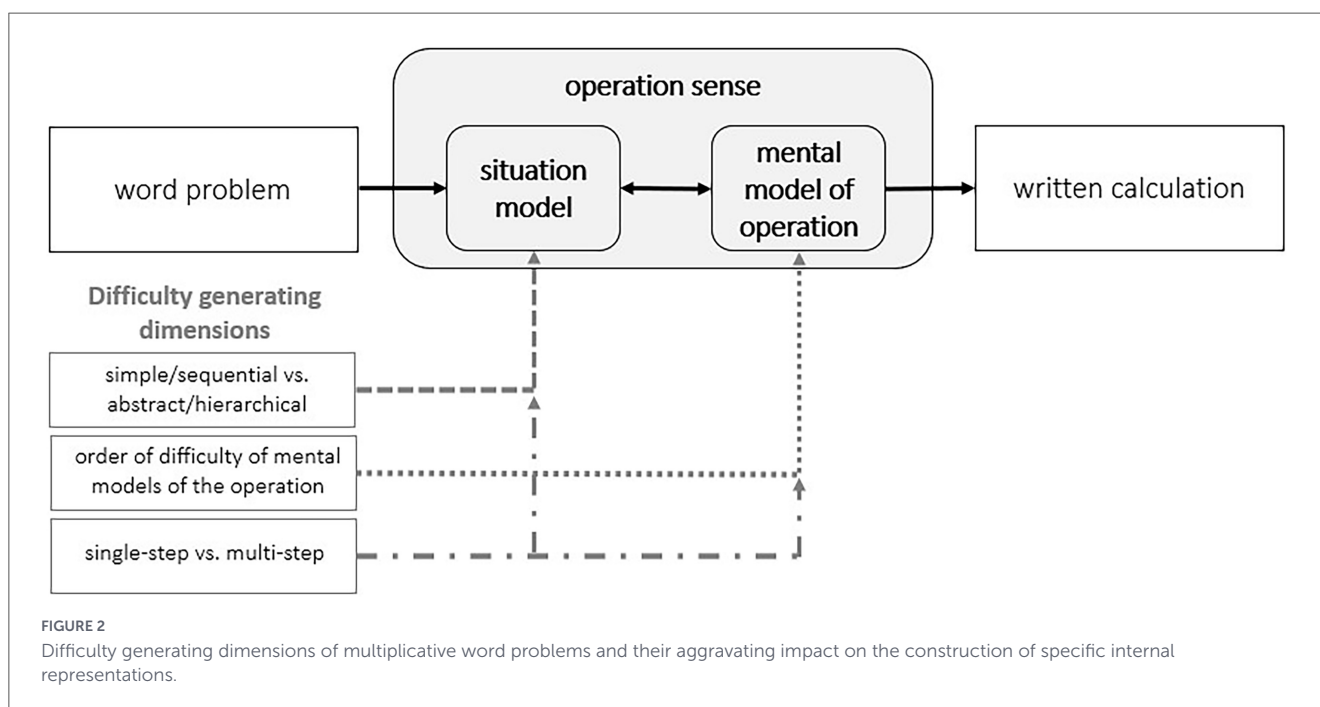
In our development, we drew on a screening instrument for basic arithmetic operation sense (Schulz et al., 2020). This screening instrument can be conducted in a class setting. It assesses operation sense in an overarching way covering all four basic operations and was developed and extensively validated based on statewide student data. Given its broad nature, covering all four basic operations within 15 items, the screening instrument by Schulz et al. (2020) is, however, not designed to distinguish differences in the multiplicative operations sense sensitively. By focusing on multiplicative operation sense only, we intend to develop a more sensitive and focused assessment of these specific skills.

In the model for the cognitive processes for solving multiplicative problems described above (cf. Figure 1), we can systematically identify crucial dimensions that contribute to

the difficulty of solving a multiplicative word problem as displayed in Figure 2. This allows us to define *levels of operation sense* based on a cognitive analysis of task features and solution processes.

In general, word problems that require a multi-step solution are assumed to be more difficult than problems with single-step solution procedures (Ehlert et al., 2013; Mayer and Hegarty, 1996; Muth, 1992). The more pieces of information that need to be processed, the harder the construction of the situation model. Moreover, in multi-step tasks, multiple mental models of operations must be activated and connected appropriately. Thus, one dimension that explains the level of difficulty of a word problem is the dimension “single-step versus multi-step word problems.” Since this dimension affects the construction of the situation model *and* the activation of the mental model of the operation, we assume that this dimension highly influences the difficulty of word problems.

There are further factors which primarily influence the construction of a viable situation model: These include the order in which information is presented (Stern, 1998) and the level of abstraction of the situation, which can be influenced e.g., by the presence or absence of keywords (Mayer and Hegarty, 1996; Verschaffel et al., 2000). If the word problem is arranged sequentially, then all information is presented in the order in which it needs to be processed. If the information is not presented in the order in which it must be processed, the arrangement is a hierarchical and mental pre-structuring is required for constructing an appropriate situation model. This additional requirement increases the level of difficulty (Franke and Ruwisch, 2010; Stern, 1998). Also, word problems which describe more abstract situations concerning relations between quantities or numbers appear to be more difficult, in particular when the keywords indicating a specific operation are missing or even misleading (Mayer and Hegarty, 1996; Verschaffel et al., 2000). Building on these findings a second dimension that explains the level of difficulty is “simple/sequential versus abstract/hierarchical problem structure.”



For multiplicative word problems there is furthermore a certain hierarchy of the underlying mental models of the operation. This means that some mental models of an arithmetic operation (definitions and example tasks for the mental models of multiplication and division are provided in [Supplementary Appendix](#)) seem to be more difficult to understand for students than others. In general, multiplication word problems seem to be easier than division word problems (Ehlert et al., 2013; Schulz et al., 2020; Schulz and Wartha, 2021). The easiest mental model of multiplication are equal groups (static/spatial-simultaneous and dynamic/temporal-successive) with direct reference to concrete action experiences (Schulz and Wartha, 2021). More difficult are multiplicative comparisons and proportions/ratios (Mulligan and Micheltore, 1997). Problems with Cartesian products appear to involve particular difficulties and can only be solved by children in later grades (Verschaffel et al., 2007).

Based on these findings, we developed the grid displayed in [Table 1](#) for classifying cognitive processes underlying multiplicative operation sense (the “cognitive model” in the framework of educational assessment, Pellegrino et al., 2001). This grid allows for categorizing word problems regarding their level of difficulty: By assigning level 1 and 2, we distinguish between one-step problems versus multi-step problems. We considered this dimension as having the most influence on task difficulty because it involves the construction of a more complex situation model *and* the activation and connection of multiple mental models of operations. Thus, in multi-step problems more cognitive processes are needed, and more mistakes can be made. Second, within each of the two levels, we distinguish between a simple, sequential structure of the situation (levels 1a and 2a) versus an abstract, hierarchical structure of the situation (levels 1b and 2b). An abstract or hierarchical structure of the situation requires mental pre-structuring to construct a viable situation model.

Building on this grid, we developed a test instrument for multiplicative operation sense with tasks on each sub-level (forming the “interpretation model” in the framework of educational assessment, Pellegrino et al., 2001). We have taken the different levels of difficulty of the specific operations into account by including the more difficult operations (e.g., division situations as deemed to be more difficult than multiplication; Ehlert et al., 2013) and the more difficult mental models of the operations (e.g., multiplicative comparisons as deemed to be more difficult than multiplicative situations on equal groups; Mulligan and Micheltore, 1997) only in tasks classified as 1b or higher, because when a word problem addresses a more difficult operation or mental model of the operation, the situation which needs to be translated also gets more abstract.

Since the Cartesian product is proven to be latest to be successfully solved arithmetically (Verschaffel et al., 2000), we considered applying the mental model of Cartesian product to require a higher level of each dimension and is thus placed in level 2b. To fulfill the dimension of the multi-step-characteristic we decided to choose Cartesian products with two factors.

The grid in [Table 1](#) leads to a four-level competence model for multiplicative operation sense, presented in more details in [Table 2](#). It is consistent with the more coarsely described model of Schulz et al. (2020).

The resulting levels require students to have different competences in order to meet these requirements of the distinct

TABLE 1 Difficulty generating dimensions as a foundation for the competence levels.

	Simple/sequential structure of the situation	Abstract/hierarchical structure of the situation
Single-step problems	1a <i>Equal groups (static/dynamic)</i>	1b <i>Multiplicative comparison quotative and partitive division</i>
Multi-step problems	2a Connection of different mental models of multiplication and division	2b <i>Cartesian product</i>

In italics: the first appearance of the required mental models of the operations.

levels: To solve items at level 1a students need to understand simple multiplicative operations in manageable situations (simplest mental model of multiplication). Items at level 1b require the understanding of multiplicative operations in more abstract situations as well as understanding word problems requiring multiplicative comparison or a division operation. In addition, students at this level must overcome superficial strategies to not get lost when the keyword might be misleading or missing.

At level 2a, the students must master sequentially constructed multi-step word problems (including all mental models of the operations multiplication and division). To solve such tasks, students must be able to link different mental models of operations sequentially (Schulz et al., 2020). For the construction of a viable situation model, all relevant elements and their relations need to be detected and kept in working memory. Sequential linking means that the linking of operations can be taken directly from the text and calculated sequentially “along the text.” Step-by-step processing of individual pieces of information is, thus, possible.

At level 2b a solution “along the text” is no longer possible. On this level students need to understand hierarchical multi-step operations and solve items with Cartesian products. Word problems on this level can only be understood by students if the hierarchical situation has been mentally pre-structured. Often, all necessary information must be considered simultaneously.

Based on this model, we developed 22 items (cf. [Supplementary Appendix](#)), which were assigned to the four levels. To minimize the complexity of the language demands and to limit the differences in item difficulty to the described dimensions, the items use simple language, syntax and grammar (e.g., mostly one-clause sentences or simple connected main sentences). The size of the numerical material was deliberately chosen in the higher number range of 11–28 to avoid solutions by automated reproduction of operation facts. We restricted the number material on natural numbers only since in our population—unlike in the curricula of other European countries—rational numbers are only taught from grade 6 onwards. Only one response format was implemented: Students are asked to write down the calculation approach (in cases of a multi-step problem the use of a placeholder is suggested). The result is explicitly not required to be calculated (the “observation” in the framework of educational assessment, Pellegrino et al., 2001).

We named our test instrument to assess the multiplicative operation sense MOvE, which is a German acronym for *Multiplicative Operationsverständnis Erfassung* (i.e., multiplicative operation sense assessment).

TABLE 2 Four level competence model of multiplicative operation sense (based on Schulz et al., 2020).

Level	Competence	Difficulty generating factors concerning the		Sample item
		Situation model	Mental model of the operation	
1a	Understanding simple single-step multiplicative operations within simple situations	Always consistent keywords. Numbers are presented in the order of their processing. Situation is directly graspable.	Multiplication: equal groups (static/dynamic)	<i>Lissi reaches into the gummy bear bag twelve times. Each time she takes 7 gummy bears. How many gummy bears does she have in total?</i>
1b	Understanding single-step multiplicative operations within more abstract/not clearly structured situations and understanding division operations	Numbers are not reliably presented in the order of their processing. Still one-step solution processes. Not necessarily consistent keywords, mental structuring of the situation and information is required.	Multiplication: comparisons and proportions/rates Division: partitive (by multiplier), quotative (by multiplicand)	<i>Tim has 100 Euros. He has five times more than Michael. How many Euros does Michael have?</i> <i>Philipp, Dilara and Jonas raised 84 euros together at the flea market. They want to share the money fairly. How much money does everyone get?</i>
2a	Understanding sequential multiplicative multi-step operations	Solution can be constructed by stepwise information processing of given information along the text.	Combination of different mental models of the operation	<i>A student company produces 160 cereal bars. These are packed 4 at a time. One pack is sold for 8 euros. How much money can the student company earn this way?</i>
2b	Understanding hierarchical multiplicative multi-step operations	Solution cannot be reconstructed by stepwise information processing of discretely considered parts of given information.	Combination of different mental models of the operation Multiplication: Cartesian product	<i>Martin threw 12 points in the basketball game. Thomas scored four times as many points. How many more points than Martin did Thomas throw?</i> <i>Sabine has 3 skirts and 4 T-shirts and 5 scarves. How many different outfits can she combine?</i>

3 The present studies

The validity of an assessment instrument is crucial when drawing inferences about students' knowledge and skills. Validity is not a fixed property of a test but an ongoing process concerning the interpretation and use of test scores (Kane, 2013; Messick, 1995). Building on the theory-driven development of an assessment instrument for multiplicative operation sense, as previously described (following Pellegrino et al., 2001), this study seeks to provide empirical evidence for its validity. The developed test is intended to allow inferences about the degree of multiplicative operation sense, and therefore, in two studies, we examine whether empirical findings support the theoretical predictions of the competence model regarding item difficulties.

In the pilot study (PS), a smaller sample was drawn to pilot the developed test items, identify potential problems, and test our main assumptions. The main study (MS) aimed to examine the stability of these assumptions while addressing the shortcomings of the pilot study. To strengthen the validity argument of the test and its underlying competence model, we addressed the following research questions in both studies:

RQ1 (PS and MS): is the empirical scaling with respect to item difficulty aligned with the theoretical prediction according to the model of multiplicative operation sense?

Our hypothesis is that the empirical item difficulty of items on level 1a is lower than that of items on level 1b, 2a, and 2b. Accordingly the empirical item difficulty of items on level 1b is lower than that of items on level 2a and 2b and the empirical item difficulty of levels of 2a is lower than items on level 2b.

RQ2 (PS and MS): to what extent do the a priori defined, theoretically grounded competence levels account for the variance in item difficulty?

Our hypothesis is that variance between the items can be explained to a large extent by their a priori defined competence level. We hypothesize that a substantial portion of the variance in item difficulty can be explained by the items' predefined competence levels.

RQ3 (PS and MS): are the levels of the model of multiplicative operation sense significantly distinct and clearly hierarchical ordered as predicted?

We conceptualized the primary dimension of item difficulty as the distinction between single-step and multi-step problems, which defined levels 1 and 2. At a subordinate level, within each of these two categories, we further differentiated item difficulty along a second dimension: the distinction between sequential and abstract problems. This created a hierarchical structure of difficulty levels. In short, our hypothesis on the item difficulty of each level is: $1a < 1b < 2a < 2b$. These assumptions appear plausible with respect to the theory and findings described above, but of course, it must be put to an empirical test.

In the main study we additionally investigated:

RQ4 (MS): are reading comprehension and basic arithmetic operation sense predictors for students' multiplicative operations sense?

Since word problems are presented in text form, and the first step in our model involves constructing a viable situation model through accurate reading, we expect reading comprehension to predict performance on the MOvE assessment. Furthermore, because our test measures multiplicative operation sense, which builds on basic arithmetic operation sense (including addition and subtraction), we also expect this mathematical prior knowledge to be a predictor of test performance.

4 Pilot study: methods and materials

4.1 Participants and procedure

66 German 5th graders participated in this study. In Germany, 5th graders usually are 10–11 years old. Multiplication and division are taught from the beginning of 2nd grade in primary school and should—regarding to the curriculum—be well understood by the end of primary school. Therefore, it can be assumed that all children had sufficient learning opportunities to acquire the basic mental models related to multiplication and division with natural numbers. Note that rational numbers in our population are only taught from grade 6th onwards.

The data was collected in a whole-class paper-pencil setting. Participants were class-wise assigned to the test versions 1 and 2. Test versions 1 and 2 contained the same items, but in different order (see below). The test was conducted by the mathematics teachers of each class, who were instructed by the first author and received an instruction manual with specific wording instructions. The students were asked to write down their calculation approach for each test item (without the need to calculate the result).

4.2 Material: MOvE - assessment of multiplicative operation sense

Two difficulty-generating dimensions for multiplicative word problems were derived from theories of text comprehension and research on word problems, resulting in a two-by-two grid (Table 1). This grid served as the basis for a four-level competence model (Table 2), for which items were designed for each theoretically derived level. To examine whether the developed items could be reliably assigned to these levels, two rounds of expert validation were conducted. In the first round, a validation manual with a decision tree was provided, outlining the theoretical and empirical foundations of the test construction. Two experts in mathematics education and educational psychology (authors 2 and 4) independently assigned each of the 23 items to one of the levels. Expert validation showed agreement on 20 of the 23 items. The random-adjusted Cohen's Kappa agreement measure was $k = 0.67$, which can be regarded as substantial agreement (Landis and Koch, 1977). The three items without agreement were discussed and finally agreed on a common assignment. In addition to this validation, a group discussion with six experts in mathematics education was conducted after which one item was excluded (because of using inappropriate easy number material) and two others were assigned to different levels (Level 1a: 5 items; Level 1b: 6 items; Level 2a: 4

items; Level 2b: 7 items) for the following reasons: When the keyword is very clear even a multiplicative comparison can be on level 1a (item 1a.5) and even without a clear keyword a situation can be imagined as a temporal-successive situation (item 1a.4).

All 22 items are fully described in [Supplementary Appendix](#). The items were distributed across two test booklets, administered in four waves, ranging from easier tasks (level 1a) to more difficult tasks (level 2b). To control item order effects (Nagy et al., 2016) and to ensure that each item was attempted by a balanced number of students, the order of items was varied across the two booklets.

To rate an item as solved the right written calculation approach (even without the calculation of the result) was enough.

4.3 Data and statistical analyses

All responses were scored dichotomously (1 = correct, 0 = incorrect) based on the correctness of the solution approach. Correctly solved items were those for which a complete and correct mathematical approach was provided. The correctness of the numerical result was considered irrelevant, as the test focused on conceptual rather than procedural skills. Correct results without a written calculation were also coded as correct, assuming that the student had applied the appropriate basic operations.

The test was designed for completion within one 45-min lesson. For each student, we identified the last item started. Missing values before this item were coded as incorrect, assuming that these tasks were considered but not solved. Missing values after the last attempted item were coded as missing, assuming that they were not started due to time restrictions.

About 50% of the data was coded by a second rater using a coding manual. The degree of agreement was determined using Cohen's kappa. With $k = 0.926$ almost perfect intercoder reliability was obtained for the whole test. Substantial intercoder reliability was obtained with $k \geq 0.716$ for each item (Landis and Koch, 1977).

There are two types of dependencies in our dataset: on the one hand, the same students answer different items, and, on the other hand, the same items are answered by different students. While only one of these two dependencies can be handled in a classical model, generalized linear mixed models (GLMM) can account for variance in both dimensions. GLMMs allow explicit tests of theoretically derived fixed effects (e.g., levels of the competence model), model comparisons between a theory-informed model and a baseline null model, and the decomposition of variance at both the person and item level. This directly tests the alignment between empirical item difficulty and the model's theoretical level structure, strengthening the evidence for construct validity. For a detailed discussion about the advantages of GLMMs compared to other statistical methods see Brauer and Curtin (2018).

To examine how well the variance in item difficulty could be explained by the theoretical levels, we applied generalized linear mixed models. The full model included fixed effects for the predictor variable *level*, corresponding to the *a priori* assigned theoretical levels 1a, 1b, 2a, and 2b of the competence model, and allowed for random intercepts for students and items.

We report both marginal and conditional R^2_{GLMM} (Nakagawa and Schielzeth, 2013) as estimates for variance explained by the fixed effects only [i.e., $R^2_{GLMM(m)}$] and the entire model including

random effects [i.e., $R^2_{GLMM}(c)$]. Effect sizes are given as odds ratios (ORs) which represent the relative increase (or decrease) in the estimated solution probability for one item if the corresponding predictor increases by 1. In addition, we report the *proportion change in variance (PCV)* on the item random intercept (Nakagawa and Schielzeth, 2013) as estimates for the variance explained by unique fixed effects, i.e., how the inclusion of specific predictors changes variance on the item random intercept. More specifically, we would regard the results of the study in line with our hypotheses, if integrating a fixed effect for *level* as a predictor would show a significant relation to the estimated solution probability and lower the random item intercept (implying a shift in variance from “assumed random” to the fixed effect representing the theoretical level the item belongs to). All analyses were conducted in *R* (R Development Core Team, 2008) using the *lme4* package (Bates et al., 2015). Plots were generated using the packages *sjPlot* (Lüdtke, 2018) and *ggplot2* (Wickham, 2016).

To evaluate if the sample size was at all big enough to confirm our hypotheses, we conducted two *post hoc* power analyses for the full model. We therefore report *post hoc* power analyses for those effects between item levels, calculated via Monte Carlo Simulation with the *simr* package (Green and MacLeod, 2016).

5 Pilot study: Results

The easiest item reached a solution probability of 89.4% (item1a.4), the hardest item a solution probability of 0.01% (item2b.5). Estimated solution probabilities with 95% CI for all single items and their assignment to the levels 1a to 2b are shown in Figure 3.

The maximum number of correct answers per student was 20, the minimum was 0. There was a high number of missing values (351 of 1452, 25%). In this pilot study missing values were not included in the analysis. Since we were interested in item characteristics not person

characteristics, we were interested in meaningful processing of the items and not in items that could not be completed due to time limit (s. the description of this procedure on in section 4.1 Materials).

RQ1: is the empirical scaling with respect to item difficulty aligned with the theoretical prediction according to the model of multiplicative operation sense?

Items that are easier to solve have a higher estimated solution rate. Thus, the estimated solution rate can be interpreted as the empirical item difficulty. If the items are scaled according to the model of multiplicative operation sense, then the data should reveal that the estimated solution rate of all items of level 1a is higher than the solution rate of all items of levels 1b, 2a, and 2b. All items of level 1b should have a solution rate lower than all items on level 1a, but higher than all items on levels 2a and 2b, etc. Thus, our hypothesis is that the empirical item difficulty of items of level 1a is lower than that of items of level 1b, 2a, and 2b—and for the other levels accordingly.

The analysis of the data shows a good dispersion of all levels beginning with an estimated solution rate of 89.4% and ending with 0.01%. So almost the whole possible span is covered (cf. Figure 3). Most of the items are scaled according to the model of multiplicative operation sense, with a decreasing solution rate with increasing level.

But results indicate that there are three items which seem to be empirically less difficult as their theoretical categorization would suggest. These items are item2a.1, item2a.2 and item2b.1. Possible reasons for this misfit, resulting in the unexpected ordering, will be discussed.

RQ2: to what extent do the *a priori* defined, theoretically grounded competence levels account for the variance in item difficulty?

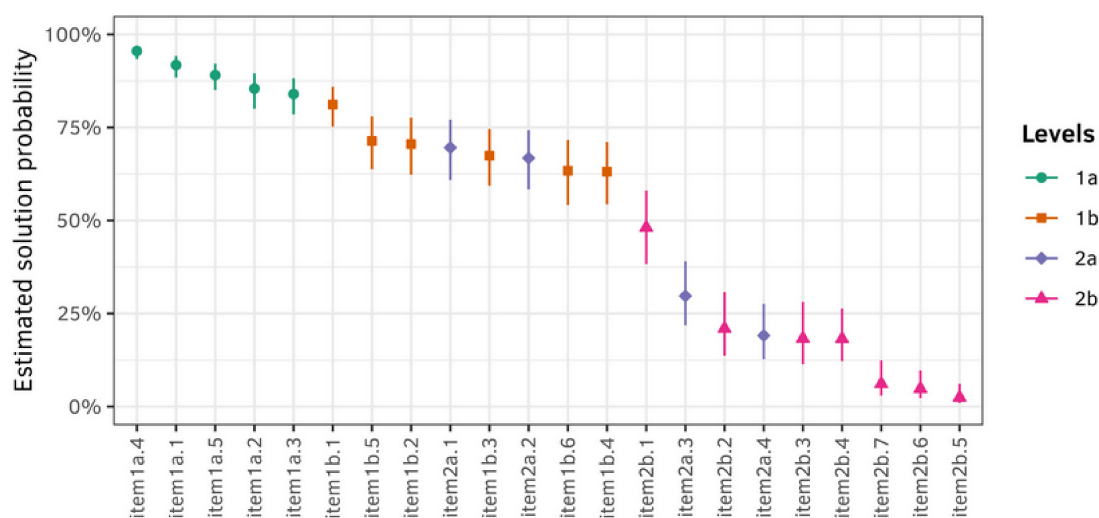


FIGURE 3
Estimated solution probability for each item with 95% CI.

TABLE 3 Parameter estimates for the generalized linear mixed models predicting the item difficulty by level assignment.

Fixed effects	Null model		Full model		Null model (slopes)		Extended model	
	OR	p	OR	p	OR	p	OR	p
Intercept	0.39		4.21	**	0.40	*	4.21	**
Level 1b			0.22	**			0.22	**
Level 2a			0.07	***			0.07	***
Level 2b			0.01	***			0.01	***
Random effects	Var.		Var.	PCV			Var.	PCV
Student	2.107		2.0877		1.94		2.05	
Item	4.037		0.5686	86%	3.58		0.58	83%
Model fit indices	Index		Index		Index		Index	
R ² GLMM(m)	0		0.333		0		0.314	
R ² GLMM(c)	0.651		0.631		NA		0.629	
AIC	1069.9		1041.3		1082.6		1056.2	
BIC	1085.0		1071.3		1142.7		1131.2	

1,101 observations, 66 students, 22 items; Pairwise between level comparisons are given as *post hoc* Tukey estimated marginal means. Estimates are given as odds ratios. PCV, Proportion change in variance on the student random intercept; R² GLMM, Marginal and Conditional estimates for variance explained (see Nakagawa and Schielzeth, 2013). Levels of significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Value in bold indicates the PCV on the item random intercept between the null model and the full model.

To answer this research question, three GLMMs were compared. The *null model* included only student ability and item difficulty as random effects. The *full model* (considering the levels) included the levels of the competence model as a categorical fixed effect. Furthermore, an *extended model* additionally included the student performance as random slopes, testing potential differential effects. This extended model was applied to test whether the complexity levels induce difficulty to the same extent for each student, e.g., to rule out that level 1a is more difficult to some students than 2a, even if the overall effect of item level is in line with the hypothesis. Our hypothesis is that variance between the items can be explained to a large extent by their *a priori* defined competence level.

A Likelihood ratio test between the *null model* (AIC = 1069.9) and the *full model* (AIC = 1041.3) showed that the *full model* better fitted the data than the *null model* which did not consider the levels of the competence model, $\chi^2(3, N = 66) = 34.62$, $p < 0.001$. Moreover, the proportion change in variance (PCV) on the item random intercept between the *null model* and the level considering *full model* was 86%, revealing that the four corresponding competence levels (to which the items were *a priori* assigned to) substantially explain the estimated item difficulty. Table 3 presents all estimates in detail.

Finally, the *extended model* reveals a slightly lower proportion change in variance in comparison with the *null model (slope)* on the item level and on the student level an even higher variance. Thus, the best fitting and most economic model remains the *full model* and the effects of the levels on the item difficulty are for all (the high and the low performing) students the same, which is an indicator for a fair test instrument.

Affirming our RQ2, the level considering full model fits the data best: the main part (86%) of the variance of the item difficulty is explained by the levels.

RQ3: are the levels of the model of multiplicative operation sense significantly distinct and clearly hierarchical ordered as predicted?

If the levels are ordered as predicted above, the mean solution probability of the items of one level would decrease with increasing level. To prove that they are significantly distinct the pairwise differences between all levels should turn significant.

Our hypothesis is that when item dimensions that increase complexity are combined with the main dimension single-step vs. multi-step, items are increasingly more difficult: $1a < 1b < 2a < 2b$.

A look at the odds ratios of the full model in Table 3 shows the following: All levels compared to the base level of level 1a are less than 1, showing that the chance of not solving the item correctly increases with increasing levels. To express it positively by considering the inverse of the odds ratio, one can state that items on level 1b are 4.5 times more difficult to solve than tasks on level 1a, $p < 0.05$, *post hoc* power of 88,6%, 95%CI [86.47, 90.50]. Tasks at level 2a are even 14-times harder than tasks at level 1a, $p < 0.001$, *post hoc* power of 99,9%, 95%CI (99.44, 100.00) And tasks at level 2b are almost 100-times harder as tasks at level 1a, $p < 0.001$, *post hoc* power of 100%, 95%CI (99.63, 100.00), almost 21-times harder than tasks at level 1b, $p < 0.001$, *post hoc* power of 100%, 95% CI (99.63; 100.00), and almost 7-times harder than tasks at level 2a, $p < 0.05$, *post hoc* power of 95.9%, 95% CI (94.48, 97.04). In short: All pairwise different effects that are significant show a *post hoc* power of over 85%. As shown in Table 3, all reported effects are significant at $p < 0.05$.

Post hoc Tukey tests showed that (in the full model that includes the levels as predictors) the estimated mean solution probabilities differ significantly between item levels ($p < 0.05$), except from stage 1b to 2a ($p = 0.165$). Confirming our hypothesis, the estimated solution rate decreased with increasing item level (level 1a: 80,8%, level 1b: 48,4%, level 2a: 23,5%, level 2b: 4,3%) as displayed in Table 4.

TABLE 4 Mean solution probabilities of the levels and pairwise differences of the levels.

Mean solution probability			Pairwise differences							
			Level 1a		Level 1b		Level 2a		Level 2b	
	Probability	SE	OR	SE	OR	SE	OR	SE	OR	SE
Level 1a	0.8081	0.0647	–	–	4.49*	2.27	13.68***	7.81	93.4***	51.05
Level 1b	0.4843	0.0958	–	–	–	–	3.05	1.65	20.82***	10.61
Level 2a	0.2354	0.0831	–	–	–	–	–	–	6.83*	3.84
Level 2b	0.0431	0.175	–	–	–	–	–	–	–	–

Mean Solution Probability = estimated marginal means; Pairwise between level comparisons are given as *post hoc* Tukey contrasts in odds ratios. SE, Standard error. Levels of significance: * $p < 0.05$, *** $p < 0.001$. *P*-value adjustment: Tukey method for comparing a family of 4 estimates, tests are performed on the log odds ratio scale.

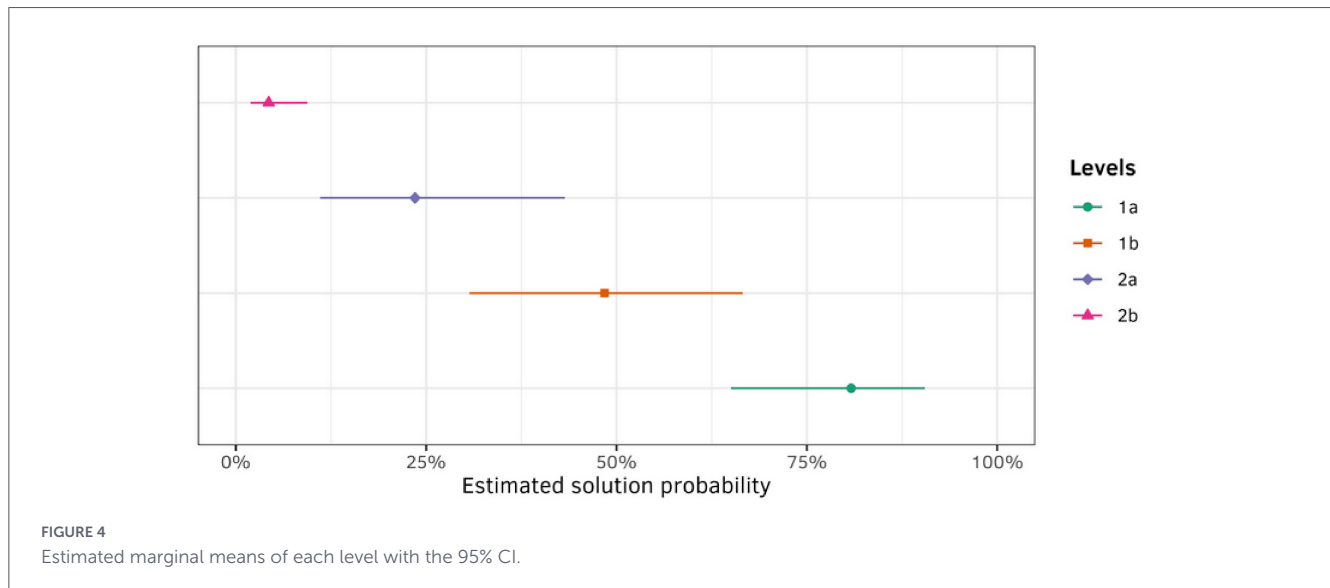


Figure 4 shows the estimated marginal means with their 95% CI showing that the mean solution probability of each level differs clearly as assumed decreasing from each level to the next higher. One can see that the 95% CI of level 1b (0.3069, 0.6657) overlaps at both ends on the upside with the 95% CI of 2a (0.1107, 0.4321) strongly and on the downside slightly with the 95% CI of level 1a (0.6503, 0.9051), which is discussed below.

We can partly affirm our RQ3 stating that the levels are distinct and ordered in the predicted way, even though there is a problem with the distinction of levels 1b and 2a. Potential causes and remedies will be discussed in the next chapter.

6 Pilot study: summary and discussion of the limitations of the pilot study

Overall, the pilot study supported the hypotheses and provided evidence for the competence model and test items, while also highlighting shortcomings to be addressed.

Three items (2a.1, 2a.2, 2b.1) emerged as outliers, showing higher solution rates than predicted. Items 2a.1 and 2a.2 may have been easier because they combined multiplicative with additive or subtractive mental models of operations; as addition and subtraction are generally less demanding (Ehlert et al., 2013; Schulz

et al., 2020), these multi-step items were easier than expected. This suggests that difficulty arises not from linking operations *per se* but from linking two demanding operations (e.g., multiplication and division). To align with their intended level, items 2a.1 and 2a.2 should be revised to combine multiplication and division only. On level 2b this facilitating impact of less complex mental models of operations did not occur. Therefore, on level 2b mental models of all operations and their combinations (with at least one multiplicative structure) are included. The main difference of difficulty on level 2b seems to be the multi-stepness combined with a hierarchical order and not the arithmetic operations which are applied.

Item 2b.1 likely appeared too easy due to simple number material (15:3 + 15), which enabled automated reproduction and guessing. In revision, more difficult number material should be used.

An overlap in item difficulty and the non-significant difference in mean solution probabilities between levels 1b and 2a should be noted. One possible explanation is that items reflecting only one difficulty-generating dimension (i.e., levels 1b and 2a; $1a < 1b = 2a < 2b$) are of similar difficulty, suggesting that the individual dimensions contribute equally to word problem difficulty. Alternatively, the small sample may have been insufficient to detect this difference. Nevertheless, meaningful results were obtained, supported by the advantages of generalized linear mixed models. A *post hoc* power analysis confirmed adequate

power for all significant effects. In the main study, differentiation between levels 1b and 2a should be improved, for example by increasing the complexity of multi-step problems through linking only demanding operations (multiplication and division), rather than including addition or subtraction.

The absence of higher-achieving students from the academic track is a limitation of the pilot study. However, as students' abilities explained only a small proportion of variance in item difficulty (see Table 3, extended model), their absence is unlikely to systematically bias the results or threaten validity. A further limitation is the lack of control for reading comprehension, a factor known to influence word problem solving (Stephany, 2021; Vilenius-Tuohimaa et al., 2008) which should be addressed in the main study.

7 Main study: methods and materials

In the main study the three critical items were revised, and a larger sample was recruited, to increase the distinction of levels and to have a broader, more representative population, that includes all strands of secondary education. Moreover, we assessed reading comprehension and basic arithmetic operation sense as prior knowledge to investigate their potential predicting effects for the test performance in the MOvE instrument.

7.1 Participants and procedure

464 German fifth graders from 22 classes of 11 secondary schools (5 *Werkrealschulen*, 5 *Realschulen*, 12 *Gesamtschulen*) took part in the present study. According to curriculum all students had already been taught the mathematical content dealt with in our study. The mean age was 11,29 years (SD 0.525); approximately 43% were female.

Concerning language background (operationalized by family language), the study involved a diverse group of participants, including both monolingual (46%) and multilingual learners (54%).

The students attended either the middle (*Realschule*) or the lowest (*Werkrealschule*) track of Germany's tripartite secondary school system or visited schools in which all students of the whole competence spectrum learn together (*Gesamtschule*). These different types of schools ensured a broad diversity in the sample, particularly concerning mathematical competence and reading comprehension. In German curricula, whole number multiplication and division instruction occurs in grades 2 until 4. Unlike other countries multiplication with rational numbers is taught only from grade 6 onwards.

The Ministry of Education responsible approved the study, and the students and their parents gave informed consent.

The data was collected in a whole-class paper pencil setting. All students worked on the test of multiplicative operation sense (MOvE) (see Supplementary Appendix) and completed a questionnaire concerning demographics (age, sex, multilingual background). The test was carried out by the mathematics teacher, who received a manual of implementation instructions. Reading ability and basic arithmetic operation sense was separately assessed

with two scales of *Lernstand 5*, a general obligatory assessment in grade 5 in our part of Germany.

7.2 Material

7.2.1 MOvE: assessment of multiplicative operation sense

The items that did not fit in in the pilot study (item2a.1, item2a.2 and item2b.1), were (accordingly to our hypothesis explained in the discussion above) replaced by new ones that dispensed with other operations and focused solely on multiplication and division (item2a.1, item2a.2) or respectively revised using more difficult number material (item2b.1). In item1a.5 the antelope was changed into a deer. All other items were retained (see Supplementary Appendix). All 22 test items were listed in the test booklet in order of their expected difficulty starting with the easiest one.

All solutions to the items were scored dichotomously (1 = correct, 0 = incorrect) regarding the correctness of the solution approach (see above). As the items were listed in the order of their expected difficulty, all missing values were coded as incorrect.

About 24% of the data was coded by a second rater using a coding manual. The degree of agreement was determined using Cohen's kappa. With $k = 0.945$ almost perfect intercoder reliability was obtained for the whole test. With $k \geq 0.661$ at least substantial intercoder reliability was obtained for each item, with almost perfect agreement ($k > 0.846$) for 20 out of 22 items (Landis and Koch, 1977).

7.2.2 *Lernstand 5*: math and reading covariate

To account for individual differences in prior knowledge and reading comprehension skills, we used the outcomes of two scales from the mandatory screening at the beginning of secondary school in Grade 5 in Baden-Wuerttemberg named "*Lernstand 5*" (Schulz et al., 2017, 2020; Fischer et al., 2017). It is assessed in the first weeks of the new school year by the teachers of Mathematics and German using standardized test materials provided by the Central Institute for Educational Analyses (IBBW) and administered uniformly to all students.

The scale for basic arithmetic operation sense assesses understanding of the four operations across 15 items in which situations (presented as texts or pictures) must be translated into adequate basic arithmetic operations. The outcome of this assessment is a suitable indicator of the students' prior knowledge, as the tasks in our intervention also required the translation of a situation into arithmetic operations (Schulz et al., 2017).

The reading comprehension scale includes 35 tasks across four texts targeting different reading skills. The tasks are designed to assess different reading skills, such as the comprehension of texts and the interpretation of information (Fischer et al., 2017).

We used the outcome measures obtained by the Central Institute for Educational Analyses (IBBW) as indicators for prior mathematical knowledge and reading comprehension, respectively.

7.3 Data and statistical analyses

For RQ1–RQ3, analysis was identical to the analysis in the pilot study.

To examine how test outcomes (MOvE) depended on learners' reading comprehension and basic operation sense, we applied generalized linear mixed models (GLMMs) comparing three models. All models included random intercepts for students and items. The null model contained only random effects (student ability, item difficulty). The full model added fixed effects for predictor variable levels, and the covariate model added fixed effects for prior knowledge (basic operational sense) and reading comprehension. All predictors were mean-centered and z-standardized to yield interpretable intercepts. Analyses were conducted in R (R Development Core Team, 2008) with the *lme4* package (Bates et al., 2015).

8 Main study: results

The easiest item reached a solution rate of 87% (item1), the hardest item a solution rate of 3% (item20) (s. Figure 5). Estimated solution probabilities with 95% CI for all single items and their assignment to the levels 1a to 2b are shown in Figure 5. The lowest achieved sum score was 0, the highest 22. The mean sum score was 9.4 with a standard deviation of 4.8. The mean of prior knowledge was 48.205 (SD 21.987) and reading comprehension 46.198 (SD 22.217). The minimum of prior knowledge was 0 and the maximum was 100; the minimum of reading comprehension was 0 the maximum was 97.

For RQ 4 (prior knowledge and reading comprehension as predictors), we excluded missing values case wise. Unfortunately, there were higher values of missings (reading comprehension: 62 missings; prior knowledge: 53 missing), because two classes did not return the list of Lernstand 5 results, which minimized the sample

for this investigation to $N = 367$. The mean in this sample for prior knowledge was 47.455 (SD 21.398) and for reading comprehension 47.000 (SD 22.089). The minimum of prior knowledge was 0 and the maximum was 100; the minimum of reading comprehension was 0 the maximum was 97.

RQ1: is the empirical scaling with respect to item difficulty aligned with the theoretical prediction according to the model of multiplicative operation sense?

Investigating RQ1 with the new sample, the analysis shows again a very good dispersion, beginning with an estimated solution rate of 93.9% and ending with 1%. Thus, almost the whole possible span is covered (cf. Figure 5). In line with our hypothesis, the empirical scaling with respect to item difficulty aligns completely with the theoretical prediction according to the model of multiplicative operation sense.

RQ2: to what extent do the *a priori* defined, theoretically grounded competence levels account for the variance in item difficulty?

To answer the research questions, we compared again three GLMMs. The *null model* included only student ability and item difficulty as random effects. The *full model* (considering the levels) included the levels of the competence model as a categorical fixed effect.

A Likelihood ratio test between the *null model* (AIC = 8393.4) and the *full model* (AIC = 8340.0) showed that the full model better fitted the data than the *null model*, $\chi^2 = (3, N = 464) = 59.42$, $***p < 0.001$. (All estimates are shown in Table 5).

Finally, the extended model reveals a slightly higher proportion change in variance in comparison with the null model (slopes) on the item level and on the student level an even higher value in variance indicating that the variance has increased rather than decreased compared to the original null model (slope). Thus, the best fitting and most economic model remains the full model and the effects of the levels on the item difficulty are for all (the high and the low performing) students the same, which is an indicator for a fair test instrument. Affirming our RQ2, the level considering full

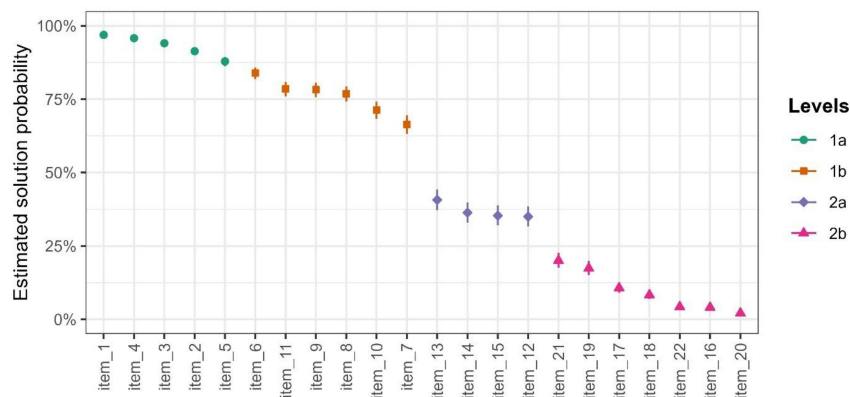


FIGURE 5
Estimated solution probability for each item with 95% CI.

TABLE 5 Parameter estimates for the generalized linear mixed models predicting the item difficulty by level assignment.

Fixed effects	Null model		Full model		Null model (slopes)		Extended model	
	OR	p	OR	p	OR	p	OR	p
Intercept	0.5		8.59	***	0.5		14.46	***
Level 1b			0.19	***			0.12	***
Level 2a			0.03	***			0.01	***
Level 2b			0.00	***			0.00	***
Random effects	Var.		Var.	PCV	Var.		Var.	PCV
Student	3.53		3.52		6.77		6.90	
Item	4.97		0.30	94%	5.37		0.27	95%
Model fit indices	Index		Index		Index		Index	
R ² GLMM(m)	0		0.391		0		0.377	
R ² GLMM(c)	0.721		0.718		0.787		0.756	
AIC	8393.4		8340.0		8253.5		8195.6	
BIC	8415.1		8383.4		8340.4		8304.1	

10,208 observations, 464 students, 22 items; Pairwise between level comparisons are given as *post hoc* Tukey estimated marginal means. Estimates are given as odds ratios. PCV, Proportion change in variance on the student random intercept; R² GLMM, Marginal and Conditional estimates for variance explained (see Nakagawa and Schielzeth, 2013). Levels of significance: ****p* < 0.001. Values in bold indicates the PCV on the item random intercept between the null model and the full model.

TABLE 6 Mean solution probabilities of the levels and pairwise differences of the levels.

	Mean solution probability				Pairwise differences					
			Level 1a		Level 1b		Level 2a		Level 2b	
	Probability	SE	OR	SE	OR	SE	OR	SE	OR	SE
Level 1a	0.8955	0.0251	-	-	5.23***	1.78	31.95***	12.10	249.3***	84.90
Level 1b	0.6213	0.0575	-	-	-	-	6.11***	2.21	47.68***	15.20
Level 2a	0.2116	0.0489	-	-	-	-	-	-	7.80***	2.78
Level 2b	0.0333	0.0077	-	-	-	-	-	-	-	-

Mean Solution Probability = estimated marginal means; Pairwise between level comparisons are given as *post hoc* Tukey contrasts in odds ratios. SE, Standard error. Levels of significance: ****p* < 0.001. *P*-value adjustment: Tukey method for comparing a family of 4 estimates, tests are performed on the log odds ratio scale.

model fits the data the best and clarifies the main part (94%) of the variance of the item difficulty is explained by the levels.

We now can affirm our RQ3 stating that the levels are significantly distinct and ordered in the predicted way.

RQ3: are the levels of the model of multiplicative operation sense significantly distinct and clearly hierarchical ordered as predicted?

RQ4: are reading comprehension and prior knowledge predictors for students' multiplicative operations sense?

A look at the odds ratios of the full model in Table 5 shows the following: All levels compared to the base level of level 1a are less than 1, showing that the chance of not solving the item correctly increases with increasing levels: Items on level 1b are 5-times more difficult to solve than tasks on level 1a. Tasks at level 2a are even 20-times harder than tasks at level 1a. And tasks at level 2b are 100 times harder. As shown in Table 5, all reported effects are significant at *p* < 0.05. *Post hoc* Tukey tests showed that (at the level considering full model) the estimated mean solution probabilities differ significantly between all item levels. All estimates are shown in Table 6.

To be able to draw conclusions about the moderating effects of basic arithmetic operation sense as prior knowledge and reading comprehension the data is analyzed using GLMMs and comparing the full model (including the levels as predictors as fixed effects) and the covariate model (including prior knowledge and reading comprehensions as predictors as fixed effects). Results show that prior knowledge and reading comprehension have a predicting effect (*p* < 0.001). Table 7 shows the models with their parameters. A one standard deviation increase in prior knowledge was associated with 2.83 higher odds of solving an item [OR = 2.83, 95% CI (2.36, 3.38)]. And a one standard deviation increase in reading comprehension was linked with 1.38 higher odds for solving an item [OR = 1.38, 95% CI (1.16, 1.65)].

Figure 6 shows the estimated marginal means with their 95% CI showing that the mean solution probability of each level differs clearly as assumed decreasing from each level to the next higher.

The PCV on the student level is in the covariate model 44%. That means that 44% of the variance of the student-level

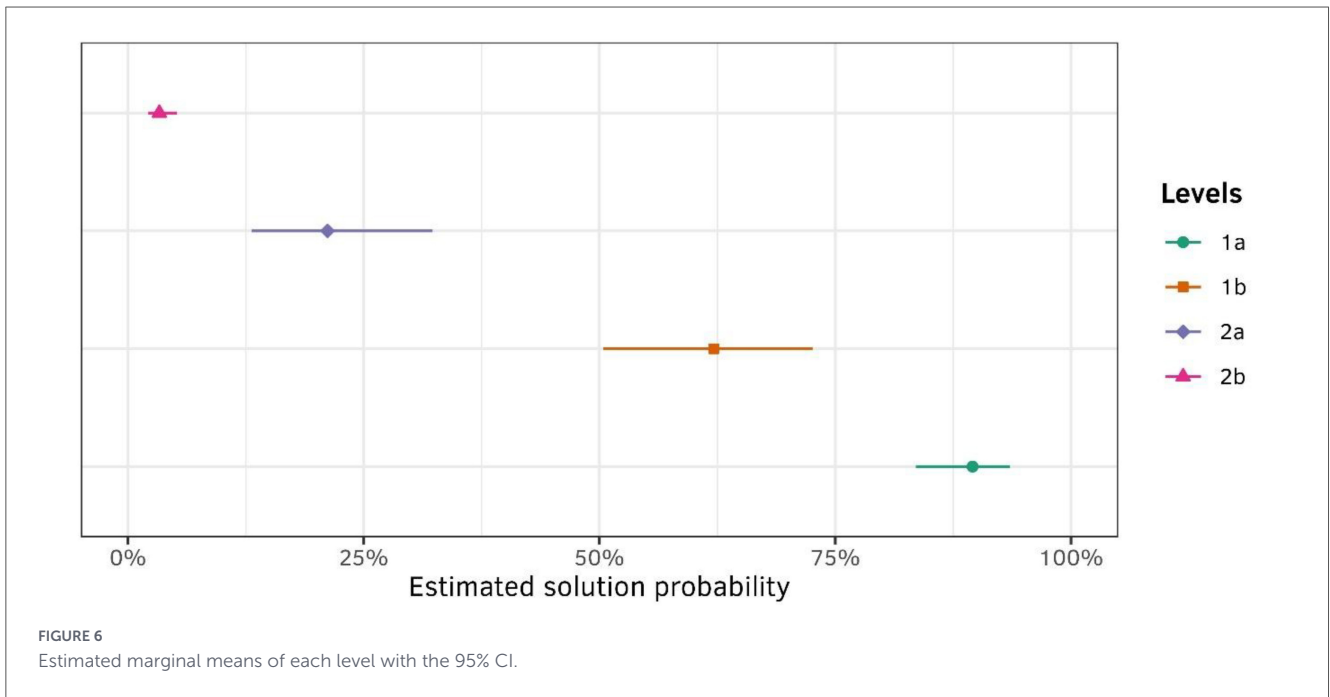


TABLE 7 Parameter estimates for the generalized linear mixed models predicting the impact of prior knowledge and reading comprehension on test performance.

Fixed effects	Null model		Full model		Covariate model	
	OR	p	OR	p	OR	p
Intercept	0.49		9.37	***	9.56	***
Level 1b			0.18	***	0.18	***
Level 2a			0.03	***	0.03	***
Level 2b			0.00	***	0.00	***
Prior knowledge					2.83	***
Reading comprehension					1.38	***
Random effects	Var.		Var.	PCV	Var.	PCV
Student	3.35		3.29		1.83	44%
Item	5.59		0.37	93%	0.37	93%
Model fit indices	Index		Index		Index	
R ² GLMM(m)	0		0.419		0.543	
R ² GLMM(c)	0.731		0.727		0.726	
AIC	6539.6		6488.9		6308.6	
BIC	6560.6		6530.9		6364.5	

8,074 observations, 367 students, 22 items; Pairwise between level comparisons are given as *post hoc* Tukey estimated marginal means. Estimates are given as odds ratios. OR, Odds ratios; Var., Variance; PCV, Proportion change in variance on the student and item random intercept; R²_{GLMM}, Marginal and Conditional estimates for variance explained (see Nakagawa and Schielzeth, 2013). Levels of significance: ****p* < 0.001. Values in bold indicate: (a) the PCV on the item random intercept between the null model and the covariate model, and (b) the PCV on the student random intercept between the null model and the covariate model.

is explained by these two covariates, whereas item-level variance remained unchanged. This indicates that the level classification explains systematic differences in item difficulty independently of student characteristics. The stability of the item-level variance further suggests that item difficulty is determined by item properties rather than by differences in the composition of the sample, providing strong evidence for construct validity.

9 Summary and discussion of both studies

In the present article, we reported about the development of a test instrument for multiplicative operation sense (MOvE) and about its empirical validation studies. By doing so, we contribute to the existing research in multiple ways:

We validated a competence model of multiplicative operation sense and the test instrument which was developed based on it. Moreover, we contributed to the broad field of research on difficulty generating factors in word problems and the influence of learner characteristics (reading comprehension and prior knowledge) on word problem solving. Based on this we derived practical implications and directions for further research

9.1 Validation of a test instrument for multiplicative operation sense

We based our development on research showing that even fifth-grade students often lack multiplicative operation sense (Brown et al., 2010; Ehlert et al., 2013; Schulz et al., 2017), which constitutes a crucial foundation for further learning in mathematics in secondary school (Hulbert et al., 2017; Schulz et al., 2017). Therefore, a testing instrument is needed to identify these deficits and provide clear starting points for targeted instructional support. A review of existing instruments (e.g., Moser Opitz et al., 2010; Booker, 2011) revealed the need for a new instrument that can be quickly applied in classrooms, assess multiplicative operation sense in a fine-grained way, and inform subsequent adaptive instructional support.

Ideally, multiplicative operation sense is developed conceptually and consolidated during the first 4 years of schooling (Gaidoschik et al., 2021). However, insufficient early support can lead to persistent deficits and intensify overtime (e.g., Gaidoschik, 2008). As mathematics is a cumulative subject, such deficits need to be addressed through appropriate support measures in lower secondary education— even if not explicitly outlined in the curriculum—before further progression in the curriculum can be considered meaningful.

Longitudinal research about how to build up and foster multiplicative operation sense effectively and long lasting from the very beginning would be beneficial.

While Schulz et al. (2020) validated a broad screening for operation sense across all arithmetic operations, it is not sensitive to different levels of multiplicative operation sense. Our aim, therefore, was to develop a more fine-grained differentiating instrument by focusing exclusively on multiplicative word problems. To this end, we refined the competence model of operation sense (Schulz et al., 2020) for multiplication and division, developed items reflecting distinct difficulty-generating factors, and demonstrated empirically that the items aligned with their assigned levels. In doing so, we followed the recommended procedure for designing and evaluating formative assessments (Pellegriano et al., 2001).

The main study addressed the limitations of the pilot study. All items now fit perfectly into the difficulty hierarchy according to their level. The mean solution probabilities of the levels differ significantly from each other. The clear hierarchical order of the levels, their significant discrimination, and the explained variance of 94% by the *a priori* theoretically explained levels clearly underpin the structural aspect of validity of the proposed competence model and test instrument for multiplicative operation sense. A scoring model can now be derived which assigns each student to a level based on their test score. The test instrument can be applied in both educational settings and research contexts.

As the test consisted of word problems, reading comprehension is essential for solving the items per design (Prediger et al., 2015). Accordingly, the strong predictive effects of reading comprehension and prior knowledge were expected and are consistent with previous findings (e.g., Pongsakdi et al., 2020; Stephany, 2021; Vilenius-Tuohimaa et al., 2008).

Notably, reading comprehension had an additional separate effect from prior knowledge, highlighting the importance of fostering both reading comprehension and basic mathematical skills in elementary school as foundational for later learning success. Our finding can, thus, be interpreted as a call for action for policy to intensify initiatives in this direction and for research to follow a differential-psychological approach to word problem research.

Gymnasium (academic track of the German tripartite school system) students were not part of the sample. But academic high-achieving students were represented as students of *Gesamtschulen* (9.2% top-level reading comprehension; 11.2% top-level basic arithmetic operation sense). Our test targets the general lower secondary level (*Sekundarstufe I*). Including *Gymnasium* students could have caused ceiling effects and thereby compromising the differentiating power of the test, as their strong performance in German and Mathematics would likely allow them to solve most items. Exclusion ensured that item analyses reflected the target population, supporting valid inferences about students' multiplicative operation sense.

9.2 Contribution to research of difficulty generating features of multiplicative word problems

Moreover, our empirical results confirm and challenge some findings about difficulty generating features of multiplicative word problems: The results support the hierarchy of the mental models of multiplication and division (Ehlert et al., 2013; Schulz et al., 2020) showing that mental models of division (item7, item8, item9) are harder than mental models of multiplication (item1, item2, item3, item4). The cartesian product (item21; item22) showed in our test to be the most difficult mental model of multiplication (Verschaffel et al., 2007). Differences in the difficulty between quotative and partitive division—like other studies report (e.g., Ehlert et al., 2014)—were not investigated. In contrast to other studies (e.g., Ehlert et al., 2013; Mayer and Hegarty, 1996), our pilot study hinted that multi-step problems when information processing along the text is possible are not harder *per se* than single-step problems but that the difficulty of the embodied mental model of the operation seems to be decisive.

The inclusion of operations beyond multiplication and division in Level 2b, given our focus on multiplicative reasoning, requires further justification: Research indicates that by the beginning of fifth grade, addition and subtraction are typically well mastered (e.g., Schulz et al., 2017), so their inclusion is unlikely to introduce additional difficulty, as confirmed by our results. This pattern suggests that, for Level 2b, the items' multi-step nature and hierarchical structure were sufficient to generate level-adequate challenge, independent of the specific operations involved.

There are more difficulty generating features in word problem construction (e.g., irrelevant information, language complexity),

TABLE 8 Scoring model for the test of multiplicative operation sense (MOvE).

Sum score	Items on level 1a	Items on level 1b	Items on level 2a	Items on level 2b	Level assignment
1–4	Partly proficient	Not proficient	Not proficient	Not proficient	1a
5–10	Proficient	Partly proficient	Not proficient	Not proficient	1b
11–14	Proficient	Proficient	Partly proficient	Not proficient	2a
15–22	Proficient	Proficient	Proficient	Partly proficient	2b

which were not addressed in our model (for an overview see Daroczy et al., 2015). We omitted irrelevant information and tried to keep language complexity low and comparable across all levels. However, there are linguistic differences in the description of multiplicative relationships when the complexity of the situation increased from simple to abstract on level 1a and 1b. Linguistic complexity may have an additional—in our study not examined—effect on item difficulty (Daroczy et al., 2015). To investigate potential effects more precisely a systematic manipulation of such variables (e.g., partitive vs quotative items; language features like passive constructions; implicit number material) would be promising.

9.3 MOvE as formative assessment

As a practical implication, the test instrument MOvE can be applied to test students in the sense of formative assessment. Based on the results of the study (scaling of items and levels aligned with the theoretically grounded model), a scoring model (Table 8) can be developed. According to Pellegrino et al. (2001) a scoring model supports the interpretation of observations in an educational assessment.

The assignment of total test sum scores to proficiency levels is based on the number of items associated with each level. For example, a student who scores 8 points has likely solved the basic multiplication tasks correctly (Level 1a: Items 1–5) and is therefore considered *proficient* at Level 1a. However, this student appears to have encountered difficulties with the more complex mental models required for multiplication and division and not solved all items on level 1b correctly (Level 1b: Items 6–11). Consequently, the student is classified as *partly proficient* at Level 1b and is therefore assigned to Level 1b. The assigned levels give information about the competence profile of the students, which can be used for targeted support that aligns with the learning needs at the corresponding level as follows: Students at levels 1a and 1b have considerable difficulties in activating appropriate mental models of multiplication, and students at level 1b struggle with the more complex mental models of multiplication and of division. For these learners, targeted support aimed at developing such mental models is crucial (e.g., using external analog representations such as arrays, e.g., investigated by Bajwa et al., 2023; Malola, 2020; Abraham and Prediger, 2025). Learners at levels 2a and 2b, in contrast, tend to experience difficulties with multi-step word problems. One explanation is that they do not build a viable situation model that contains all important elements and relations. Building heuristic strategies that promote focusing on the relevant quantities and their relations is especially necessary for these students (e.g., with concept maps as graphic scaffolds like suggested by Dröse and Prediger, 2021).

10 Conclusion

The competence model of multiplicative operation sense provides a framework for advancing formative assessment and research. The presented test instrument allows reliable and valid measurement of individual differences and can be used to evaluate adaptive support interventions and their differential effects (Leuders et al., 2020). The focus on multiplication and division — operations that have been shown to be particularly problematic at the beginning of secondary school (Brown et al., 2010; Ehlert et al., 2013) — together with the possibility of a more precise assessment of these operations that can be directly implemented in practice to guide adaptive support, constitutes a key strength of the presented test. Future work should further investigate cognitive processes in solving multiplicative word problems and conduct intervention studies to test remedial approaches in ecologically valid settings.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical approval was not required for the study involving human samples in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

ES: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing, Project administration, Resources, Supervision, Validation, Visualization. TL: Conceptualization, Methodology, Supervision, Validation, Writing – review & editing. FR: Formal analysis, Methodology, Visualization, Writing – review & editing. KL: Conceptualization, Methodology, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This research was conducted within the Research Training Group “HeLPS – Heterogeneity: effective learning settings and professionalism in schools,” funded internally by the University of Education Freiburg, Baden-Wuerttemberg.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

References

- Abraham, M., and Prediger, S. (2025). Scaffolding fifth graders' learning with a digital multi-representation applet: Design research on focusing multiplicative structures with dynamic dot arrays. *Dig. Exp. Math. Educ.* 11, 219–246. doi: 10.1007/s40751-024-00156-7
- Anderson, R. C. (2018). “Role of the reader's schema in comprehension, learning, and memory,” in *Theoretical Models and Processes of Literacy*, eds D. E. Alvermann, N. J. Unrau, and R. B. Ruddell (London: Routledge), 136–145. doi: 10.4324/9781315110592-9
- Bajwa, N. P., Tobias, J. M., and Lawton, C. (2023). Children's conceptions on the structure of an array: Using quick images as a gateway to multiplicative ideas. *J. Math. Behav.* 69:101049. doi: 10.1016/j.jmathb.2023.101049
- Baroody, A. J., Lai, M., and Mix, K. S. (1999). “The Development of young children's early number and operation sense and its implications for early childhood education,” in *Handbook of Research on the Education of Young Children*, eds B. Spodek and O. N. Saracho (Mahwah, NJ: Lawrence Erlbaum Associates Publishers).
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Booker, G. (2011). *Building Numeracy: Moving from Diagnosis to Intervention*. Melbourne: Oxford University Press.
- Borromeo Ferri, R. (2006). Theoretical and empirical differentiations of phases in the modelling process. *Zentralblatt für Didaktik der Mathematik* 38, 86–95. doi: 10.1007/BF02655883
- Brauer, M., and Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychol. Methods* 23, 389–411. doi: 10.1037/met0000159
- Brown, M., Kuchemann, D., and Hodgen, J. (2010). “The struggle to achieve multiplicative reasoning,” in *Proceedings of the British Congress of Mathematics Education (BCME7)*, ed. A. Joubert (Manchester: University of Manchester), 49–56.
- Daroczy, G., Wolska, M., Meurers, W. D., and Nuerk, H. C. (2015). Word problems: A review of linguistic and numerical factors contributing to their difficulty. *Front. Psychol.* 6:348. doi: 10.3389/fpsyg.2015.00348
- Dröse, J., and Prediger, S. (2021). Identifying obstacles is not enough for everybody—differential efficacy of an intervention fostering fifth graders' comprehension for word problems. *Stud. Educ. Eval.* 68:100953. doi: 10.1016/j.stueduc.2020.100953
- Ehlert, A., Fritz, A., Arndt, D., and Leutner, D. (2013). Arithmetische Basiskompetenzen von Schülerinnen und Schülern in den Klassen 5 bis 7 der Sekundarstufe. *J. Math. Didakt.* 34, 237–263. German. doi: 10.1007/s13138-013-0055-0
- Ehlert, A., Wolf, A., Kess, J., and Fritz, A. (2014). Schülerkompetenzen zum Dividieren beim Übergang zwischen Primar- und Sekundarstufe. *Empirische Pädagogik* 28, 319–337.
- Fischbein, E., Deri, M., Nello, M. S., and Marino, M. S. (1985). The role of implicit models in solving verbal problems in multiplication and division. *J. Res. Math. Educ.* 16, 3–17. doi: 10.5951/jresmetheduc.16.1.0003
- Fischer, U., Merz, G., and Wagner, S. (2017). Kompetenzorientierung im Leseverstehensunterricht. Verknüpfung von Diagnose und Förderung in Lernstand 5. *Leseforum Schweiz*. doi: 10.58098/lfl/2017/3/610
- Franke, M., and Ruwisch, S. (2010). *Didaktik des Sachrechnens in der Grundschule*, 2nd Edn. Heidelberg: Springer Spektrum. German.
- Gaidoschik, M. (2008). Rechen Schwäche in der Sekundarstufe: Was tun? *J. Mathe. Didakt.* 29, 287–294. doi: 10.1007/BF03339065 German.
- Gaidoschik, M., Opitz, E. M., Nührenbörger, M., and Rathgeb-Schnierer, E. (2021). *Besondere Schwierigkeiten beim Mathematiklernen. Leitlinie der Gesellschaft für Didaktik der Mathematik. Mitteilungen der Gesellschaft für Didaktik der Mathematik, (111S)*, 3–19. German. Available online at: <https://ojs.didaktik-der-mathematik.de/index.php?journal=mgdm&page=article&op=view&path%5B%5D=1042> (accessed June 22, 2023).
- Gravemeijer, K. (1997). Solving word problems: A case of modelling? *Learn. Instr.* 7, 389–397. doi: 10.1016/S0959-4752(97)00011-X
- Green, P., and MacLeod, C. J. (2016). simr: An R package for power analysis of generalized linear mixed models by simulation. *Methods Ecol. Evol.* 7, 493–498. doi: 10.1111/2041-210X.12504
- Greer, B. (1994). “Extending the meaning of multiplication and division,” in *Multiplicative Reasoning in the Learning of Mathematics*, eds G. Harel and J. Confrey (Albany, NY: State University of New York Press), 61–88.
- Greer, B. (1997). Modelling reality in mathematics classrooms: The case of word problems. *Learn. Instr.* 7, 293–307. doi: 10.1016/S0959-4752(97)00006-6
- Gröhlich, C., Scharenberg, K., and Bos, W. (2009). Wirkt sich Leistungsheterogenität in Schulklassen auf den individuellen Lernerfolg aus? *J. Bildungsforsch.* 1, 86–105. German. doi: 10.25656/01:4557
- Hackenberg, A. J., and Tillema, E. S. (2009). Students' whole number multiplicative concepts: A critical constructive resource for fraction composition schemes. *J. Math. Behav.* 28, 1–18. doi: 10.1016/j.jmathb.2009.04.004
- Hulbert, E. T., Petit, M. M., Ebby, C. B., Cunningham, E. P., and Laird, R. E. (2017). *Focus on Multiplication and Division: Bringing Research to the Classroom*. New York, NY: Routledge.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/feduc.2026.1756297/full#supplementary-material>

- Jitendra, A. K. (2019). "Using schema-based instruction to improve students' mathematical word problem solving performance," in *International Handbook of Mathematical Learning Difficulties: From the Laboratory to the Classroom*, eds A. Fritz, V. G. Haase, and P. Räsänen (Berlin: Springer), 595–609.
- Jitendra, A. K., Dougherty, B., Sanchez, V., Harwell, M. R., and Harbour, S. (2023). Building conceptual understanding of multiplicative reasoning content in third graders struggling to learn mathematics: A feasibility study. *Learn. Disabil. Res. Pract.* 38, 285–295. doi: 10.1111/ldrp.12322
- Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, MA: Harvard University Press.
- Kane, M. (2013). The argument-based approach to validation. *School Psych. Rev.* 42, 448–457. doi: 10.1080/02796015.2013.12087465
- Kasper, D., Köller, O., Selzer, C., Wendt, H., Schwippert, K., McElvany, N., et al. (2020). *TIMSS 2019: Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*, 1st Edn. New York, NY: Waxmann. German.
- Kintsch, W., and Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychol. Rev.* 109–129. doi: 10.1037/0033-295X.92.1.109
- Kosko, K. W. (2019). A multiplicative reasoning assessment for fourth and fifth grade students. *Stud. Educ. Eval.* 60, 32–42. doi: 10.1016/j.stueduc.2018.11.003
- Lamon, S. J. (2006). *Teaching fractions and ratios for understanding. Essential content knowledge and instructional strategies for teachers. 2. Aufl.* Mahwah, NJ: Lawrence Erlbaum Associates.
- Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310
- Leuders, T., Loibl, K., and Weigand, G. (2020). Differenzierungsstrategien auf den Ebenen Lernen, Unterricht und Schule – Forschungsstände und Forschungsansätze," in eds M. Kampshoff and C. Wiepke *Vielfalt in Schule und Unterricht: Konzepte und Debatten im Zeichen der Heterogenität*, 38–52 (Stuttgart: Kohlhamer). German.
- Lüdecke, D. (2018). *sjPlot: Data visualization for statistics in social science: R package version 2.1*.
- Malola, M. (2020). The use of arrays in the learning of multiplication word problems in primary school. *Afr. Educ. Res. J.* 8, 432–441. doi: 10.30918/AERJ.82.20.033
- Marshall, S. P. (1995). *Schemas in Problem Solving*. New York, NY: Cambridge University Press.
- Mayer, R. E., and Hegarty, M. (1996). "The process of understanding mathematical problems," in *The Nature of Mathematical Thinking*, eds R. J. Sternberg and T. Ben-Zeev (Mahwah, NJ: L. Erlbaum Associates), 29–53.
- McVee, M. B., Dunsmore, K., and Gavelek, J. R. (2005). Schema theory revisited. *Rev. Educ. Res.* 75, 531–566. doi: 10.3102/00346543075004531
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.* 50, 741–749. doi: 10.1037/0003-066X.50.9.741
- Ministry of Education (2008). *Numeracy Professional Development Projects Book 2: The Diagnostic Interview*. Wellington, NZ: Ministry of Education.
- Moser Opitz, E., Reusser, L., Moeri Müller, M., Anliker, B., Wittich, C., and Freseemann, O. (2010). *BASIS-MATH 4-8*. Bern: Hans Huber Verlag.
- Moss, J., and London McNab, S. (2011). "An approach to geometric and numeric patterning that fosters second grade students' reasoning and generalizing about functions and co-variation," in *Early Algebraization*, eds J. Cai and E. Knuth (Berlin: Springer), 277–301. doi: 10.1007/978-3-642-17735-4_16
- Mulligan, J., and Michelmore, M. (1997). Young children's intuitive models of multiplication and division. *J. Res. Math. Educ.* 28, 309–330. doi: 10.2307/749783
- Mullis, I. V. S., Martin, M. O., Foy, P., and Arora, A. (2012). *Timss 2011 International Results in Mathematics*. Chestnut Hill, MA: International Association for the Evaluation of Educational Achievement (IEA).
- Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., and Fischbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science*. Boston, MA: TIMSS & PIRLS International Study Center.
- Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., and Preuschoff, C. (2009). *TIMSS 2011 Assessment Frameworks*. Boston, MA: TIMSS & PIRLS International Study Center.
- Muth, K. D. (1992). Extraneous information and extra steps in arithmetic word problems. *Contemp. Educ. Psychol.* 17, 278–285. doi: 10.1016/0361-476X(92)90066-8
- Nagy, G., Lüdtke, O., and Köller, O. (2016). Modeling test context effects in longitudinal achievement data: Examining position effects in the longitudinal German PISA 2012 assessment. *Psychol. Test Assessm. Model.* 58, 641–670. doi: 10.25656/01:12804
- Nakagawa, S., and Schielzeth, H. (2013). A general and simple method for obtaining R² from generalized linear mixed-effects models. *Methods Ecol. Evol.* 4, 133–142. doi: 10.1111/j.2041-210x.2012.00261.x
- Nesher, P., and Hershkovitz, S. (1994). The role of schemes in two-step problems: Analysis and research findings. *Educ. Stud. Math.* 26, 1–23. doi: 10.1007/BF01273298
- Pellegrino, J. W., Chudowsky, N., and Glaser, R. (eds) (2001). *Knowing what Students Know: The Science and Design of Educational Assessment*. Washington, DC: National Academies Press.
- Piaget, J. (1976). *Die Äquilibration der kognitiven Strukturen*. Stuttgart: Ernst Klett. German.
- Pongsakdi, N., Kajamies, A., Veermans, K., Lertola, K., Vauras, M., and Lehtinen, E. (2020). What makes mathematical word problem solving challenging? Exploring the roles of word problem characteristics, text comprehension, and arithmetic skills. *ZDM Math. Educ.* 52, 33–44. doi: 10.1007/s11858-019-01118-9
- Prediger, S. (2008). The relevance of didactic categories for analyzing obstacles in conceptual change: Revisiting the case of multiplication of fractions. *Learn. Instr.* 18, 3–17. doi: 10.1016/j.learninstruc.2006.08.001
- Prediger, S., Wilhelm, N., Büchter, A., Gürsoy, E., and Benholz, C. (2015). Sprachkompetenz und Mathematikleistung - Empirische Untersuchung sprachlich bedingter Hürden in den Zentralen Prüfungen 10. *J. Math. Didakt* 36, 77–104. doi: 10.1007/s13138-015-0074-0
- R Development Core Team. (2008). *R: A language and environment for statistical computing*. Available online at: <http://www.R-project.org>
- Reusser, K. (1989). *Vom Text zur Situation zur Gleichung. Kognitive Simulation von Sprachverständnis und Mathematisierung beim Lösen von Textaufgaben*. Bern: Universität Bern.
- Royar, T. (2013). *Handlung - Vorstellung - Formalisierung: Entwicklung und Evaluation einer Aufgabenreihe zur Überprüfung des Operationsverständnisses für Regel- und Förderklassen*. Hamburg: Verlag Dr. Kovac. German.
- Scheibling-Sève, C., Pasquinelli, E., and Sander, E. (2020). Assessing conceptual knowledge through solving arithmetic word problems. *Educ. Stud. Math.* 103, 293–311. doi: 10.1007/s10649-020-09938-3
- Schnotz, W., and Bannert, M. (2003). Construction and interference in learning from multiple representations. *Learn. Instruct.* 13, 141–156. doi: 10.1016/S0959-4752(02)00017-8
- Schulz, A., Leuders, T., and Rangel, U. (2017). "Arithmetische Basiskompetenzen am Übergang zu Klasse 5 - eine empirie- und modellgestützte Diagnostik als Grundlage für spezifische Förderentscheidungen," in *Handbuch Rechenschwäche*, eds A. Fritz, S. Schmidt, and G. Ricken (Weinheim: Beltz), 396–416. German.
- Schulz, A., Leuders, T., and Rangel, U. (2020). The use of a diagnostic competence model about children's operation sense for criterion-referenced individual feedback in a large-scale formative assessment. *J. Psychoeduc. Assess.* 38, 426–444. doi: 10.1177/0734282918823590
- Schulz, A., and Wartha, S. (2021). *Zahlen und Operationen am Übergang Primar-/Sekundarstufe: Grundvorstellungen aufbauen, festigen, vernetzen*. Berlin: Springer. German. doi: 10.1007/978-3-662-62096-0
- Slavit, D. (1998). The role of operation sense in transitions from arithmetic to algebraic thought. *Educ. Stud. Math.* 37, 251–274. doi: 10.1023/A:1003602322232
- Sowder, J., Armstrong, B., Lamon, S., Simon, M., Sowder, L., and Thompson, A. (1998). Educating teachers to teach multiplicative structures in the middle grades. *J. Math. Teach. Educ.* 1, 127–155. doi: 10.1023/A:1009980419975
- Stephany, S. (2021). "The influence of reading comprehension on solving mathematical word problems: A situation model approach," in *Diversity Dimensions in Mathematics and Language Learning*, eds A. Fritz, E. Gärsoy, and M. Herzog (Berlin: De Gruyter), 370–395. doi: 10.1515/9783110661941-019
- Stern, E. (1998). *Die Entwicklung des mathematischen Verständnisses im Kindesalter*. Lengerich: Pabst. German.
- Vergnaud, G. (1994). "Multiplicative conceptual fields: What and Why?," in *Multiplicative Reasoning in the Learning of Mathematics*, eds G. Harel and J. Confrey (Albany, NY: State University of New York Press), 41–60.
- Verschaffel, L., Greer, B., and Corte, E. D. (2000). *Making Sense of Word Problems. Contexts of learning*, Vol. 8. Lisse: Swets & Zeitlinger.
- Verschaffel, L., Greer, B., and Corte, E. D. (2007). "Whole number concepts and operations," in *Second Handbook of Research*, ed. F. Lester (Charlotte, NC: Information Age Publishing Inc), 557–628.
- Verschaffel, L., Schukajlow, S., Star, J., and van Dooren, W. (2020). Word problems in mathematics education: A survey. *ZDM Math. Educ.* 52, 1–16. doi: 10.1007/s11858-020-01130-4
- Vilenius-Tuohimaa, P. M., Aunola, K., and Nurmi, J. -E. (2008). The association between mathematical word problems and reading comprehension. *Educ. Psychol.* 28, 409–426. doi: 10.1080/01443410701708228
- Vom Hofe, R., and Blum, W. (2016). "Grundvorstellungen" as a category of subject-matter didactics. *J. Math. Didakt.* 37, 225–254. German. doi: 10.1007/s13138-016-0107-3
- Wartha, S., and Güse, M. (2009). Zum Zusammenhang zwischen Grundvorstellungen zu Bruchzahlen und arithmetischem Grundwissen. *J. Math. Didakt.* 30, 256–280. German. doi: 10.1007/BF03339082
- Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*, 2nd Edn. New York, NY: Springer.