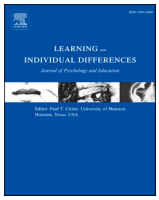
Learning and Individual Differences 114 (2024) 102518

Contents lists available at ScienceDirect

Learning and Individual Differences

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



Reality-based tasks for competency-based education: The need for an integrated analysis of subject-specific, linguistic, and contextual task features

Dominik Leiss [a,*,1], Timo Ehmke [b,2], Lena Heine [c,3]

[a] Institute of Mathematics and its Didactics, Faculty of Education, Leuphana University Lüneburg, Universitätsallee 1, 21335 Lüneburg, Germany
[b] Institute of Educational Sciences, Faculty of Education, Leuphana University Lüneburg, Universitätsallee 1, 21335 Lüneburg, Germany
[c] Institute of German Studies, Faculty of Philology, Ruhr University Bochum, GB 5/145, 44780 Bochum, Germany

ARTICLE INFO

Keywords:
Competency-based education
Assessment
Test items
Mathematical competence
Language competence

ABSTRACT

In evaluating competency-based education, effective test instruments must address real-life complexities. The impact of subject-specific, linguistic, and contextual task features, alongside central personal characteristics, on the empirical challenge of such tasks is unclear. We developed mathematics tasks from 30 real-world contexts, each with three questions of varying complexity, administered through a systematically rotated experimental design to 535 German grades 9 and 10 students. Various student variables were collected. Generalized linear mixed models revealed that contextual and mathematical task features significantly contributed to task difficulty variance. Language features had no intermediate-level influence, while students' mathematical self-efficacy moderated low task context familiarity's impact. These findings guide the construction of reality-based mathematics tasks to tailor empirical difficulty.

Educational relevance: In the context of worldwide competence-orientated education, it is crucial to reform in-class and national tests. Traditional task formats are limited in representing authentic problems. In most school subjects, a lack of understanding exists in designing reality-based competence-oriented tasks that ensure fair test conditions and meet the subjects' normative demands. This study addresses this gap by empirically investigating the interplay of subject-related, linguistic, and contextual aspects of reality-based tasks in mathematics. Teachers and researchers can use these insights to improve competence-oriented performance situations, sparking further questions. These findings encourage similar studies across subjects for broader applicability.

1. Introduction

Preparing students for real-world challenges is a fundamental goal in education (Stacey, 2015). Consequently, many countries' school education reforms in the last millennium have emphasized stronger competence orientation based on educational standards (e.g., Klieme et al., 2004; National Council of Teachers of Mathematics, 2000). Competence-oriented education imparts subject knowledge and equips students with skills to use this knowledge in authentic situations, fostering their participation as responsible citizens in society (Niss, 2003; Stacey, 2015; Weinert, 2001). In mathematics teaching, with traditionally features tasks with minimal context and limited real-world relevance (Verschaffel et al., 2000), a need for reform in task culture arises (Stein et al., 1996). Therefore, research on reality-based tasks, which have a serious contextual embedding and real-life relevance (Boaler, 2001; Den Heuvel-Panhuizen, 2003), has gained increased attention.

Consequently, the reform efforts of school tasks that had already begun (Stein et al., 1996) and related research activities have significantly intensified, for example, in the field of mathematics (Verschaffel et al., 2020; Vos, 2018). However, the findings concerning reality-based tasks are primarily limited to the learning situation, with minimal application to performance. Both in school examination formats (Drüke-Noe & Kühn, 2017; Scheja, 2017; Vos, 2013) and in international large-

* Corresponding author.
E-mail addresses: leiss@leuphana.de (D. Leiss), tehmke@leuphana.de (T. Ehmke).
[1] Research interests: mathematical literacy, language education in mathematics, competence-oriented assessments, and university teacher education.
[2] Research interests: large-scale assessments, language in the classroom, German as a second language, and university teacher education.
[3] Research interests: language education, immigrated students, multilingualism, and testing language competencies.

https://doi.org/10.1016/j.lindif.2024.102518
Received 24 August 2023; Received in revised form 12 July 2024; Accepted 17 July 2024
Available online 26 July 2024

scale assessments with competence-oriented frameworks (Mullis & Martin, 2017; OECD, 2019), primarily less complex, context-neutral, and highly linguistically reduced reality-based tasks are used owing to various framework conditions, including test duration, psychometric test fairness, or educational policy acceptance (Tout & Spithill, 2015).

Accordingly, despite the initially differentiated approaches with relatively few complex tasks (contexts) (Ferrara et al., 2011; Vondrová et al., 2019), our knowledge about reality-based tasks is still limited, particularly regarding how heterogeneous student groups are affected by various task dimensions. A knowledge gap exists concerning the interaction of reality-based task variables that have empirically been demonstrated to be significant for the solution processes (mathematical, linguistic, and contextual features), and student characteristics (mathematical, linguistic, and contextual knowledge; mathematical self-efficacy; and socioeconomic status). Therefore, we aim to experimentally examine systematically controlled mathematical reality-based task characteristics in their interaction with student characteristics. This will provide the necessary diagnostic and supportive insights into reality-based application tasks in mathematics and further (replicative) research approaches for other subjects. This is because the lack of knowledge about the differentiated descriptive difficulties in realistic tasks is a desideratum that exists in many school subjects (El Masri et al., 2017).

2. Theoretical background

Following the objective of preparing students for real life, it is crucial in the school context to address the phenomena that can be encountered in the real world (Gravemeijer et al., 2017). In this framework, the school's situation of confronting the real-world problem is authentic to a limited extent. Hence, to address relevant aspects of a real-life topic, the problems considered must not be mere brief descriptions of the situation clothed in real facts (Niss et al., 2007). Rather, the aim is to address questions relevant to real-world contexts through the application of mathematical models. The processing of these problems, referred to as reality-based tasks below, involves numerous cognitive processes—differing from classic internal mathematical tasks. The term "modeling cycle" is established to represent such solution processes (Depaepe et al., 2015). While nuances in modeling cycles may exist among subjects and within the mathematical discipline (Komor et al., 2021; Perrenet & Zwaneveld, 2012; Verschaffel et al., 2000), the expanded model in Fig. 1 (based on Blum and Leiss (2007) incorporates elements proven empirically relevant. According to Giere (1999), this is a representational conception of a model, accordingly representing an idealized description of mathematical solution processes of reality-based tasks. This is influenced by a complex interplay of personal and task-related factors, described in detail in Section 2.1 (central influencing points in italics).

(1) By successfully reading the task (personal: *language competence*), students recognize the textually transmitted (task: *number of characters* and *linguistic complexity*) real-world situation the task is embedded in. Against the background of contextually relevant experiences (personal: *socioeconomic status* and *contextual prior knowledge*/task: *familiarity of the context*), students form a situation model as an individual mental representation of the task consisting of an understanding of the context and the relevant question. (2) They then need to identify the relevant information from the text for a successful solution of the mathematical task (personal: *language competence*/task: *total number of data*). This requires the initially more complex situation model to be reduced to the so-called real model, containing only the relevant factors. (3) Depending on the complexity of the question (task: *cognitive level* and *number of procedures*), this real model is transferred into a mathematical model (e.g., a linear equation) of a specific mathematical subject area (personal: *intra-mathematical competence*/task: *curricular position*). (4) Based on this, an intra-mathematical solution to the problem is generated using mathematical procedures (personal: *intra-mathematical competence*), which may require a longer process of mathematical processing (personal: *mathematical self-efficacy*/task: *cognitive level* and *number of procedures*). (5) This intra-mathematical result must be interpreted regarding the received lifeworld context (personal: *intra-mathematical competence* and *contextual prior knowledge*). (6) Finally, the plausibility, appropriateness, and correctness of the result must be evaluated at various process stations, for example, based on lifeworld considerations (personal: *socioeconomic status* and *contextual prior knowledge*/task: *familiarity of the context*) or mathematical control strategies (personal: *intra-mathematical competence*). It is important to bear in mind that these six stages do not have to be passed through linearly and that all process activities are not only determined by cognitive factors but are also influenced by motivational factors (task: interestingness of the context).

Tasks requiring these six stages to be solved show a high degree of difficulty for students (Daroczy et al., 2015; De Bock et al., 2003). In their overview of the current state of research on reality-based tasks, Verschaffel et al. (2020) show that students are only rarely confronted with complex real-world situations, even in competence-oriented mathematics classes, and that research activities in this area often focus on the classroom learning situation with such tasks. Consequently, there are few empirical studies on the complex interplay of solution-relevant factors in competence-oriented performance situations with reality-based tasks. We present the central findings from these studies below.

2.1. Student characteristics

Various student characteristics influence mathematical performance

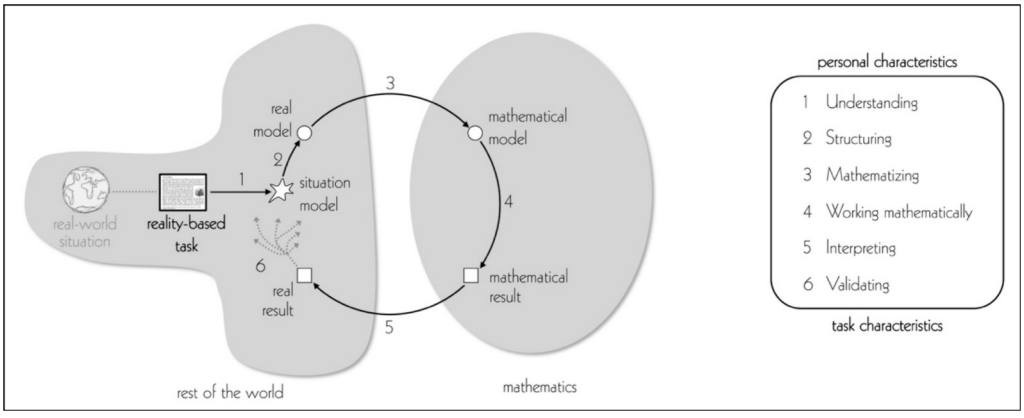


Fig. 1. Expanded modeling cycle of reality-based tasks.
(Based on Blum and Leiss (2007)).

in reality-based tasks, such as intra-mathematical competence (Berkowitz & Stern, 2018; Fuchs et al., 2006; Pongsakdi et al., 2020), language competence (Boonen et al., 2016; Peng et al., 2020; Prediger et al., 2018), cultural and socioeconomic background (Lubienski, 2000; Piel & Schuchart, 2014; Werning et al., 2008), mathematical self-efficacy (Pajares & Graham, 1999; Schunk & Pajares, 2009; Williams & Williams, 2010), and contextual prior knowledge (Ärlebäck & Bergsten, 2013; Stillman, 2000; Thevenot, 2017). However, it is largely unclear to what extent these student characteristics in connection with specific characteristics of reality-based tasks influence the probability of students solving such tasks. More extensive findings are available at least on these specific task characteristics. The limitation here, however, is that their interaction has not been adequately studied thus far, with these characteristics considered instead in isolation, so that individual findings are available on the following three task characteristics in particular: 1) mathematical, 2) linguistic, and 3) contextual.

2.1.1. Mathematical variables

Studies have investigated the various difficulty features of (reality-based) mathematics tasks. Among the most established constructs in the analysis of mathematical test items is the distinction of different *cognitive levels* in successful item completion. For instance, as one example of the over 20 existing revisions (Marzano & Kendall, 2006) of the so-called Bloom's taxonomy (Bloom et al., 1956), Anderson and Krathwohl (2001) as well as the National Assessment Governing Board (2022) distinguish three levels of complexity (low, moderate, and high), which describe the different mathematical activities: (1) recall, (2) comprehension, and (3) application. Particularly regarding the consideration of reality-based tasks, it is crucial to indicate that this distinction, successfully employed in various studies, refers exclusively to the (intra) mathematical activities, that is, in the modeling cycle; in particular, to the fourth and partly the third step. Accordingly, it can be assumed for such complex reality-based tasks that difficulty is enhanced with an increase in the cognitive level of the mathematical activities (Embretson & Daniel, 2008).

In addition to the type of cognitive process, the degree of cognitive load is an important factor. It increases with the *number of procedures* to be performed and the amount of numbers to be processed. In Cognitive Load Theory (Sweller, 2010), the limitation of working memory is central (Zheng & Gardner, 2019). As each additional procedure is typically associated with the processing of additional data, these two task characteristics are highly correlated (Raghubar et al., 2010) and can be considered one element. Owing to the reduced situation descriptions in traditional mathematics tasks, the so-called extraneous load, caused by superfluous and possibly distracting data, is often minimized (Nurjanah & Retnowati, 2018). Realistic and authentic problem situations, however, are characterized by complexity and typically entail irrelevant information, so although the students' interest in the context could increase (Hidi & Harackiewicz, 2000), the external load is also likely to increase, with an empirically ambiguous result for the task's probability of solution. Accordingly, in realistic tasks it can be assumed that the total number of data in the context and the amount of numerical data to be identified for answering the question influence the *difficulty of data selection* and thus also the correct completion of the task (Voyer, 2011).

Moreover, the subject area within which the operations are located plays an important role. Each mathematical topic has specific requirements related not only to its complexity but also to typical learner misconceptions (Russell et al., 2009); therefore, this should be considered when analyzing task difficulty (Turner et al., 2015). Empirical studies show, however, that it is not the major mathematical topics—geometry, probability, numbers, or algebra—that are associated with different levels of difficulty (Kim-O, 2011). Rather, it is the *curricular position* and thus the intensity of previous engagement with a topic that explains the difficulty of the related mathematics tasks. The empirical difficulty is affected by how long ago the topic addressed in the task was introduced or how intensively it is repeatedly addressed within a spiral curriculum. For example, Pollitt et al. (2007) indicated that the different curricular locations of a test requirement in different subject groups had a considerable influence on the solution rate (59 % vs. 79 %) in the different test groups.

Within the framework of national educational standards or large-scale assessments, typically, a single criterion (e.g., the cognitive level in TIMSS 2023 (Mullis et al., 2021)) or a more summary criterion (e.g., the combination of cognitive level and cognitive load in the German educational standards (Blum et al., 2016)) is used to describe the *general task difficulty*. Even if research projects sometimes differentiate between these characteristics (Ferrara et al., 2011; Pongsakdi et al., 2020), the number of such studies is still relatively small and the variance explanation of task difficulty remains unsatisfactory. Accordingly, El Masri et al. (2017, p. 61) conclude that previous research can only explain 20–23 % of the variance and that research is still needed to improve methods of predicting item difficulty.

2.1.2. Linguistic requirements

Forming a situation model, which is an individual's mental representation of the described situation is crucial in solving mathematical tasks (Leiss et al., 2019). Words, sentence structures, and text features are the central sources from which the building of students' mental models takes its starting point. Therefore, the linguistic requirements of a task can be assumed to play a central role in task difficulty. It is well understood for general text comprehension that linguistic characteristics can cause difficulty; however, the exact effects in subject-specific task contexts are still relatively unclear. Neri and Retelsdorf (2022) examined 40 studies on the influence of different word-, sentence-, and text-level variables on science and mathematics task performance and showed that language characteristics usually associated with difficulty generation (e.g., word frequency, sentence complexity) only inconsistently caused empirical task difficulty. Other studies that investigated linguistic simplification of mathematical text tasks failed to find an impact on the solution rate, whether in primary (Haag et al., 2015) or in secondary school (Walkington et al., 2019).

A significant limitation of most large-scale tests and related studies is the construction of a non-biased test, particularly regarding the language competencies of non-native speakers (Abedi, 2011). Thus, mathematical task texts are generally short (*number of characters*) and lexically and syntactically simple (*linguistic complexity*) (Österholm & Bergqvist, 2012). For example, the texts of sample mathematics tasks in studies such as those based on the PISA 2012 (OECD, 2012) or TIMSS 2019 (Mullis & Martin, 2017) had an average length of only 318 and 235 characters and a readability index (LIX; Björnsson, 1968) of 26 and 18, respectively, appropriate for grades 5 and 3. They were thus far below the average linguistic level that can be assumed for the tested regular students.

In contrast to the task level, relatively stable empirical findings regarding the influence of language-related personal characteristics exist. High correlations have repeatedly been found between learners' general language competencies and their ability to solve mathematical tasks (Ding & Homer, 2020). However, the extent to which students' general language competencies moderate the linguistic demands of mathematics tasks is largely unclear (Shaftel et al., 2006) and initial studies show that math-specific language skills in particular also need to be considered (Vanluydt et al., 2021). More targeted research into the interaction effects between students' (content-specific) language competences and mathematical test tasks' linguistic features is therefore warranted (Neri & Retelsdorf, 2022).

2.1.3. Context variables

"The real world is a complex and messy place, thus real-world task contexts should reflect this reality and the embeddedness of the task should […] be at least wrapper where task solvers must consider the context" is a conclusion of Brown (2019, p. 76), who analyzed the

contexts of reality-based tasks in mathematics education, in final examinations and large- scale assessment, based on a literature review. According to Smith and Morgan (2016), three central aspects can be identified as intentions for this contextualization of mathematics tasks in mathematics didactic research: the meaning of mathematics in the real world should be recognized, students should be able to engage with (mathematical) aspects of the real world, and the context should help in learning mathematics. Concerning the first two points, there is already a longer tradition in mathematics didactics research of addressing the (normative) question of what kind of reference to the real world is used or should be used (De Lange, 1996; Garrett et al., 2016; Palm, 2008; Stillman, 1998).

However, empirical findings regarding the performance-related influence of contexts are less clear. While continuous exposure to real-world contexts in mathematics seems to have a positive effect on the image of mathematics in the real world (Gijsbers et al., 2020) and also on general performance (Vos, 2020), there is still a clear lack of research on the specific influence of a concrete task context on the individual solution rate. Julie (2007) and Schukajlow et al. (2012) were able to show that students evaluate the interestingness of contexts of reality-based tasks or different task types differently. Accordingly, it must be assumed that the *interestingness of a task context* (De Bock et al., 2003; Graham et al., 2008; Pulkkinen et al., 2022; Renninger et al., 2002; Scheidemann et al., 2022) as well as *familiarity of the task context* (Albarracín et al., 2022; Walkington et al., 2015) has an influence on the probability of solving a mathematics task. The extent to which these characteristics interact with other difficulty-generating factors to explain additional task difficulty is largely unclear.

Furthermore, there are no empirical findings on whether the potential influence of *interestingness* and *familiarity of the task context* is moderated by specific personal characteristics and if so, to what extent. However, regarding the significance of external life circumstances and internal psychological factors for interest formation, it can be assumed that *socioeconomic background* (Cooper & Dunne, 1999; Halim et al., 2021), *mathematical self-efficacy* (Bong et al., 2015; Talsma et al., 2018), and *contextual prior knowledge* (Ainley et al., 2002; Stillman, 2000; Thevenot, 2017) are possible influencing factors for dealing with more or less interesting and familiar contexts.

2.2. Research questions

Based on this background, the following research questions were formulated:

1) To what extent can the empirical difficulty of reality-based tasks be influenced by systematic variations in mathematical characteristics?

2) To what extent can (a) linguistic aspects and (b) real-world context-related aspects of the task explain additional variance in empirical difficulty?

3) To what extent do student characteristics moderate the influence of different task characteristics on empirical difficulty?

3. Methods

3.1. Reality-based tasks

To address the research questions, mathematical tasks with reference to reality were specifically designed. They covered a broad spectrum of mathematical complexity and fulfilled certain requirements regarding contextual and linguistic features. Only content-related characteristics in the areas of math, language and task context were investigated. Structural aspects such as answer formats (Katz et al., 2000), item positions (Le, 2007), and test environment (Drijvers, 2018) remained constant.

3.1.1. Task contexts and linguistic characteristics

A lengthy expository text presented the context for each task based on authentically situated problems. Compared with large-scale assessments, this should target mathematical literacy much more closely. First, expository texts about real-world topics were written to include data and functional connections, but no (mathematical) questions (Fig. 2).

Thirty contexts (C01 through C30) were created covering a wide variety of topics, including wind power, Netflix series, hibernation, postal charges, jewelry production, moving walkways at airports, dentists, and YouTube. All tasks contained in the sense of Elia and Philippou (2004) one or two representational graphics, which, however, did not contain any information beyond the text and were therefore not intended to represent an influencing variable, but merely had a comparable illustrative and motivating function for all tasks (Dewolf et al., 2014). The extent to which a task was considered interesting for the pupils was recorded using the mean value of a pupil survey. For this purpose, all 30 contexts were assessed by at least 150 students (mean [M] = 240, standard deviation [SD] = 58, min = 152, max = 308) regarding the *interestingness of the context* (C_7)[4] on a 4-point Likert scale (1: not interesting, 2: slightly interesting, 3: somewhat interesting, 4: very interesting) (M = 2.2, SD = 0.97). The context-specific mean of this survey measured the *interestingness of the context*. For the assessment of the more abstract level of familiarity, all task contexts were rated by two coders on a 4-point scale (1: no personal reference, 2: potentially personal reference, 3: personal environment, 4: personal reference) regarding the *familiarity of contexts* (C_8) for students (M = 2.6, SD = 1.2). Accordingly, interestingness and familiarity with the context are two aggregated values and not individual assessments.

Compared to the various contexts addressed in the texts, the linguistic features between the 30 contexts were largely consistent: all texts were in a length range (L_5) with a mean of 2090 *characters* (SD = 236, min = 1591, max = 2737) and a similar level of *linguistic complexity* (L_6) from the LIX interval [40, 60] (M = 51.8, SD = 4.3, min = 42.1, max. = 59.0). Thus, they presented an age-appropriate reading challenge (Anderson, 1983).

3.1.2. Mathematical task characteristics

Questions (subsequently referred to as *items*) were developed only after the 30 contexts had been created. All require a mathematical analysis of contextual situations using secondary school mathematical tools and models from the field of functional relationships (the combination of a context and an item is referred to as a *task* see Fig. 3).

For each context, three differently challenging items were formulated at general task complexity (GTC) levels 1, 2, and 3, so that a total of 90 tasks (30 contexts × 3 items at three GTC levels) were available. The formulated items should, on the one hand, represent authentic questions about the context described and, on the other hand—and this can be seen as a limitation in terms of authenticity for some tasks—include subject-specific varied elements to answer the first research question. The items were assigned to the GTC levels based on the following four mathematical characteristics ($Ma_{1–4}$).

3.1.2.1. Number of procedures (Ma_1). The minimum number of arithmetic operations required was counted. An arithmetic operation describes the mathematical connection between two numerical data, resulting in a new data element. Simultaneously, this captures the amount of data needed from the context. For the constructed tasks, this characteristic covers the interval [1, 9] with an M of 2.68 and an SD of 1.67.

3.1.2.2. Curricular position (Ma_2). Curricular position plays a major role in when and how intensively a mathematical topic has been dealt

[4] The different mathematical ($M_{1–4}$), linguistic ($L_{5–6}$), and contextual ($C_{7–8}$) characteristics of the tasks as well as the different personal variables ($P_{9–13}$) are labeled by capital letters with index.

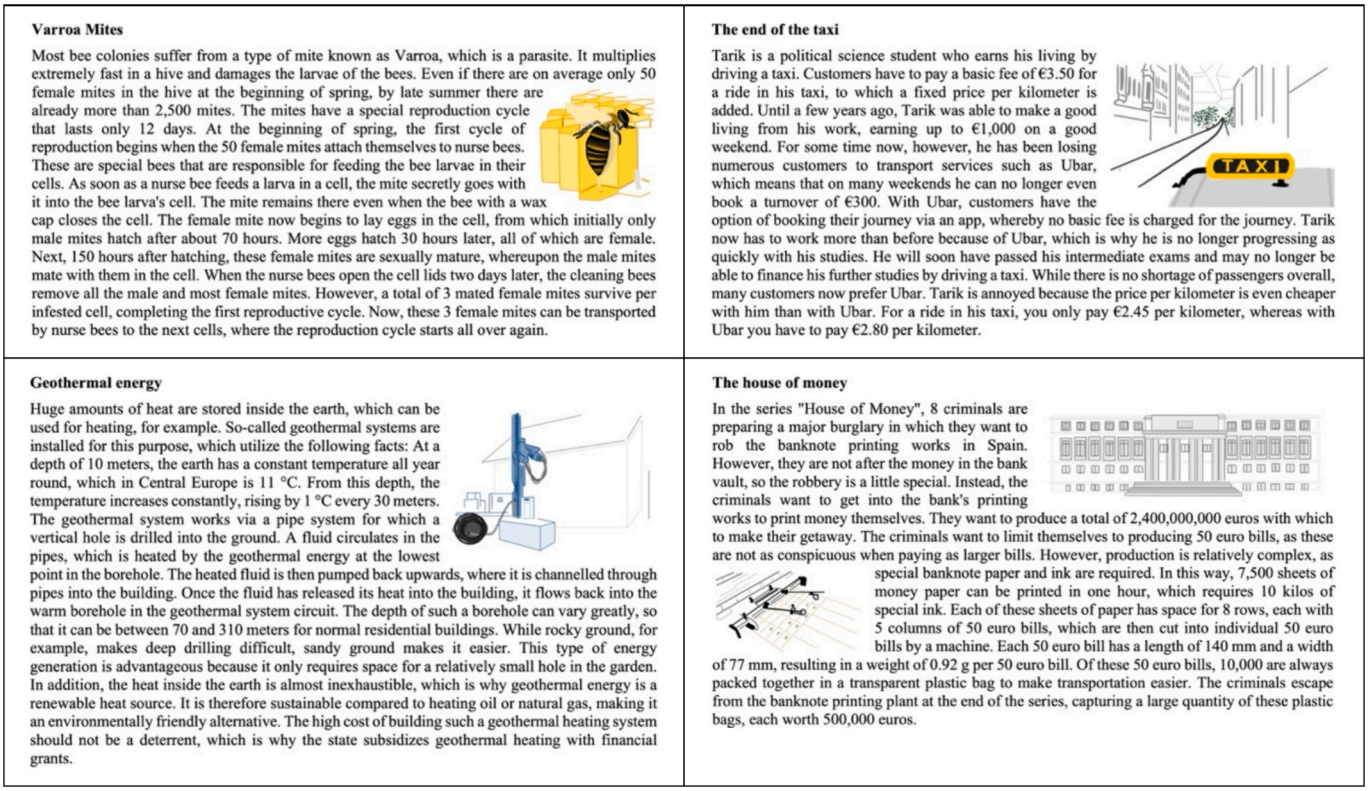


Fig. 2. Four examples of task contexts.

with. The mathematical contents addressed in the tasks were therefore evaluated regarding their curricular position as follows: Level 1, contents of the 5th/6th grade; Level 2, contents of the 7th/8th grade; Level 3, contents of the 9th/10th grade.

3.1.2.3. Difficulty of data selection (Ma_3). To mimic the central features of complex real-world problems, all tasks contained more numerical data (M = 8.09, SD = 2.05) than were necessary to answer the items (M = 3.17, SD = 1.32). To model the task-specific difficulty of data selection, we linked these two values—Number of all data (D) and Number of solution-relevant data (d)—with a hypergeometric distribution. Assuming a correct number of draws, we determined the probability of randomly drawing the solution-relevant data from the given total data of a task:

$$P(X = d) = \frac{\binom{d}{d}\binom{D-d}{d-d}}{\binom{D}{d}}$$

3.1.2.4. Cognitive level (Ma_4). This criterion only covers the quality of the (intra)mathematical activities, so that only the formation of the mathematical model and the mathematical calculations were assessed. Level 1 activities involve the direct application of basic procedures from a single mathematical subject area. At Level 2, content from several mathematical areas must be combined, and at Level 3, a complex application of various mathematical procedures is required.

In particular, the last qualitative characteristic of the cognitive level, which was also used in the present study owing to its importance in numerous (inter)national assessments (see Section 2.1.1), illustrates the necessity of double coding. Accordingly, all task characteristics were double-coded by two raters. As the rating was ordinal-scaled, inter-rater agreement was calculated using linear weighted Cohen's kappa (Cohen, 1968). κ-Values between 0.72 and 0.94 were achieved; therefore, a reliable assignment of the task characteristics was assumed.

The GTC level for each task was assigned by the maximum of the four characteristics mentioned above. For this purpose, three levels were defined for M_1 [Level 1: one operation; Level 2: 2–3 operations; Level 3: >3 operations] and M_4 [Level 1: probability <0.025; Level 2: probability between 0.025 and 0.075; Level 3: probability >0.075]. This resulted in 30 tasks with GTC level 1, 30 tasks with GTC level 2, and 30 tasks with GTC level 3 (Fig. 3).

In summary, at the first GTC, a functional relationship (e.g. tripling of mites per reproductive cycle) usually had to be recognized and the associated size calculated for a given x (50 mites * 3 = 150 mites). At the second GTC level, the same functional relationship generally had to be applied several times (50 mites go through five reproduction cycles) and at the third GTC level, several (different) functional relationships had to be linked together in context or the functional relationship had to be formulated in general terms (number of mites after x reproduction cycles). Proportional (*House of money*), linear (*The End of the Taxi* and *Geothermal Energy*), and exponential relationships (*Varroa Mites*) were addressed in the 30 contexts.

3.2. Test design

An experimental design was employed with within-item factors (mathematical, linguistic, and contextual item characteristics). Additionally, we included students' intra-mathematical skills, language proficiency, and other personal characteristics as (moderator) variables. For this purpose, a rotated paper-pencil test design with 18 different test booklets was constructed using 90 systematically varied reality-based tasks. Each test booklet consisted of the following four parts:

Part 1: Fifteen reality-based tasks (60 min).
Part 2: Intra-mathematical test (10 min).
Part 3: General language test (5 min).
Part 4: Questionnaire on student characteristics (5 min).

A total of 80 min was spent on each test booklet. The test booklet

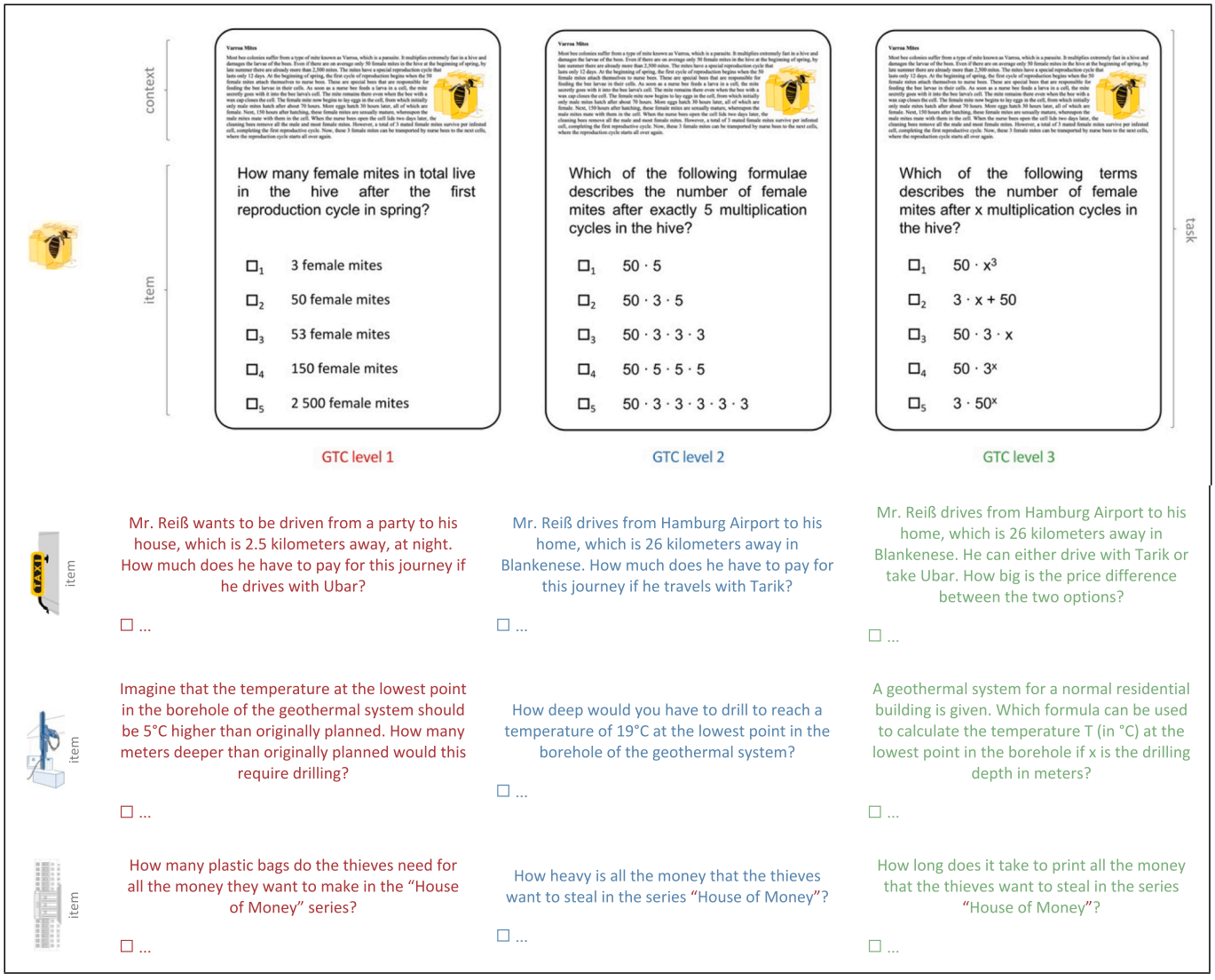


Fig. 3. Items of the four example contexts at general task complexity (GTC) levels 1–3.

versions differed only in terms of the tasks used in Part 1.

3.2.1. *Test with reality-based tasks (Part 1)*

As each student could only answer one item per context, a rotational design was necessary for Part 1 (Table 1). This ensured that the entire task pool was tested, that there were at least 50 answers for each task, that each context was used in a constant task block in terms of (mathematical) content regardless of the GTC level, and that each test booklet contained an equal number of tasks from the three GTC levels to maintain a constant theoretical level of difficulty. Each student had to work on three task blocks of five tasks in Part 1. The task blocks resulted from distributing the items for the 30 contexts (C01–C30) and the three GTC levels (1, 2, and 3) in five-task blocks across the 18 test booklets, according to the rotational pattern shown in Table 1. In addition, six reality-based tasks on functional relationships with a wide range of solution rates (35–85 %) were integrated to achieve a stronger link between the different test booklets for Rasch scaling (Fischer et al., 2021) (see Section 3.4).

3.2.2. *Intra-mathematical competence test (Part 2)*

To measure *intra-mathematical competence* (P_9), a test was constructed that consisted of 11 mathematics tasks that exclusively included content central to the mathematical idea of a *functional* relationship. These were intra-mathematical tasks without significant text with an average processing time of <1 min. One point was awarded for each correct answer, and no point was awarded for incorrect answers (M = 3.48, SD = 2.01). The test results were Rasch scaled and showed a good fit to the Rasch model with an average point-biserial correlation of 0.41, an EAP/PV reliability of 0.78 and a weighted mean square between 0.87 and 1.16. Accordingly, the wle personal scores could be used as a reliable measure of the students' internal mathematical competence.

3.2.3. *General language test (Part 3)*

Students' *language competence* (P_{10}) was tested using a modified C-Test (Eckes & Grotjahn, 2006). The test consists of a narrative text on the topic of "health in the workplace" and comprises 96 words. Twenty-five words were evenly distributed throughout the text and the second half of each word was deleted and replaced by a gap. These gaps had to be filled by the students, with 1 point awarded for each correct answer and 0 points for an incorrect answer (M = 19.2, SD = 4.81). This test, used in modified form in other studies, was also Rasch scaled and showed an average point-biserial correlation of 0.60, an EAP/PV reliability of 0.77 and a weighted mean square between 0.82 and 1.19. The wle values calculated here can therefore be regarded as reliable indicators of linguistic ability.

Table 1
Rotational design for the 18 test booklets (TB) in Part 1.

TB1	TB2	TB3	TB4	TB5	TB6	TB7	TB8	TB9	TB10	TB11	TB12	TB13	TB14	TB15	TB16	TB17	TB18
Mat01gtc1	Mat01gtc2	Mat01gtc3															
Mat02gtc3	Mat02gtc1	Mat02gtc2															
Mat03gtc2	Mat03gtc3	Mat03gtc1															
Mat04gtc1	Mat04gtc2	Mat04gtc3															
Mat05gtc3	Mat05gtc1	Mat05gtc2															
Mat06gtc2	Mat06gtc3	Mat06gtc1	Mat06gtc2	Mat06gtc3	Mat06gtc1												
Mat07gtc1	Mat07gtc2	Mat07gtc3	Mat07gtc1	Mat07gtc2	Mat07gtc3												
Mat08gtc3	Mat08gtc1	Mat08gtc2	Mat08gtc3	Mat08gtc1	Mat08gtc2												
Mat09gtc2	Mat09gtc3	Mat09gtc1	Mat09gtc2	Mat09gtc3	Mat09gtc1												
Mat10gtc1	Mat10gtc2	Mat10gtc3	Mat10gtc1	Mat10gtc2	Mat10gtc3												
Mat11gtc3	Mat11gtc1	Mat11gtc2	Mat11gtc3	Mat11gtc1	Mat11gtc2	Mat11gtc3	Mat11gtc1	Mat11gtc2									
Mat12gtc2	Mat12gtc3	Mat12gtc1	Mat12gtc2	Mat12gtc3	Mat12gtc1	Mat12gtc2	Mat12gtc3	Mat12gtc1									
Mat13gtc1	Mat13gtc2	Mat13gtc3	Mat13gtc1	Mat13gtc2	Mat13gtc3	Mat13gtc1	Mat13gtc2	Mat13gtc3									
Mat14gtc3	Mat14gtc1	Mat14gtc2	Mat14gtc3	Mat14gtc1	Mat14gtc2	Mat14gtc3	Mat14gtc1	Mat14gtc2									
Mat15gtc2	Mat15gtc3	Mat15gtc1	Mat15gtc2	Mat15gtc3	Mat15gtc1	Mat15gtc2	Mat15gtc3	Mat15gtc1									
			Mat16gtc1	Mat16gtc2	Mat16gtc3	Mat16gtc1	Mat16gtc2	Mat16gtc3	Mat16gtc1	Mat16gtc2	Mat16gtc3						
			Mat17gtc3	Mat17gtc1	Mat17gtc2	Mat17gtc3	Mat17gtc1	Mat17gtc2	Mat17gtc2	Mat17gtc3	Mat17gtc1						
			Mat18gtc2	Mat18gtc3	Mat18gtc1	Mat18gtc2	Mat18gtc3	Mat18gtc1	Mat18gtc2	Mat18gtc3	Mat18gtc1						
			Mat19gtc1	Mat19gtc2	Mat19gtc3	Mat19gtc1	Mat19gtc2	Mat19gtc3	Mat19gtc1	Mat19gtc2	Mat19gtc3						
			Mat20gtc3	Mat20gtc1	Mat20gtc2	Mat20gtc3	Mat20gtc1	Mat20gtc2	Mat20gtc3	Mat20gtc1	Mat20gtc2						
						Mat21gtc2	Mat21gtc3	Mat21gtc1	Mat21gtc2	Mat21gtc3	Mat21gtc1	Mat21gtc2	Mat21gtc3	Mat21gtc1			
						Mat22gtc1	Mat22gtc2	Mat22gtc3	Mat22gtc1	Mat22gtc2	Mat22gtc3	Mat22gtc1	Mat22gtc2	Mat22gtc3			
						Mat23gtc3	Mat23gtc1	Mat23gtc2	Mat23gtc3	Mat23gtc1	Mat23gtc2	Mat23gtc3	Mat23gtc1	Mat23gtc2			
						Mat24gtc2	Mat24gtc3	Mat24gtc1	Mat24gtc2	Mat24gtc3	Mat24gtc1	Mat24gtc2	Mat24gtc3	Mat24gtc1			
						Mat25gtc1	Mat25gtc2	Mat25gtc3	Mat25gtc1	Mat25gtc2	Mat25gtc3	Mat25gtc1	Mat25gtc2	Mat25gtc3			
									Mat26gtc3	Mat26gtc1	Mat26gtc2	Mat26gtc3	Mat26gtc1	Mat26gtc2	Mat26gtc3	Mat26gtc1	Mat26gtc2
									Mat27gtc2	Mat27gtc3	Mat27gtc1	Mat27gtc2	Mat27gtc3	Mat27gtc1	Mat27gtc2	Mat27gtc3	Mat27gtc1
									Mat28gtc1	Mat28gtc2	Mat28gtc3	Mat28gtc1	Mat28gtc2	Mat28gtc3	Mat28gtc1	Mat28gtc2	Mat28gtc3
									Mat29gtc3	Mat29gtc1	Mat29gtc2	Mat29gtc3	Mat29gtc1	Mat29gtc2	Mat29gtc3	Mat29gtc1	Mat29gtc2
									Mat30gtc2	Mat30gtc3	Mat30gtc1	Mat30gtc2	Mat30gtc3	Mat30gtc1	Mat30gtc2	Mat30gtc3	Mat30gtc1
												Mat1gtc1	Mat1gtc2	Mat1gtc3	Mat1gtc1	Mat1gtc2	Mat1gtc3
												Mat2gtc3	Mat2gtc1	Mat2gtc2	Mat2gtc3	Mat2gtc1	Mat2gtc2
												Mat3gtc2	Mat3gtc3	Mat3gtc1	Mat3gtc2	Mat3gtc3	Mat3gtc1
												Mat04gtc1	Mat04gtc2	Mat04gtc3	Mat04gtc1	Mat04gtc2	Mat04gtc3
												Mat05gtc3	Mat05gtc1	Mat05gtc2	Mat05gtc3	Mat05gtc1	Mat05gtc2
															Mat06gtc2	Mat06gtc3	Mat06gtc1
															Mat07gtc1	Mat07gtc2	Mat07gtc3
															Mat08gtc3	Mat08gtc1	Mat08gtc2
															Mat09gtc2	Mat09gtc3	Mat09gtc1
															Mat10gtc1	Mat10gtc2	Mat10gtc3
anchor task1	anchor task1	anchor task1	anchor task1	anchor task1	anchor task1	anchor task1	anchor task1	anchor task1	anchor task1	anchor task1	anchor task1	anchor task1	anchor task1	anchor task1	anchor task1	anchor task1	anchor task1
anchor task2	anchor task2	anchor task2	anchor task2	anchor task2	anchor task2	anchor task2	anchor task2	anchor task2	anchor task2	anchor task2	anchor task2	anchor task2	anchor task2	anchor task2	anchor task2	anchor task2	anchor task2
anchor task3	anchor task3	anchor task3	anchor task3	anchor task3	anchor task3	anchor task3	anchor task3	anchor task3	anchor task3	anchor task3	anchor task3	anchor task3	anchor task3	anchor task3	anchor task3	anchor task3	anchor task3
anchor task4	anchor task4	anchor task4	anchor task4	anchor task4	anchor task4	anchor task4	anchor task4	anchor task4	anchor task4	anchor task4	anchor task4	anchor task4	anchor task4	anchor task4	anchor task4	anchor task4	anchor task4
anchor task5	anchor task5	anchor task5	anchor task5	anchor task5	anchor task5	anchor task5	anchor task5	anchor task5	anchor task5	anchor task5	anchor task5	anchor task5	anchor task5	anchor task5	anchor task5	anchor task5	anchor task5
anchor task6	anchor task6	anchor task6	anchor task6	anchor task6	anchor task6	anchor task6	anchor task6	anchor task6	anchor task6	anchor task6	anchor task6	anchor task6	anchor task6	anchor task6	anchor task6	anchor task6	anchor task6

3.2.4. Questionnaire on students' characteristics (part 4)

To investigate potential moderator effects, three variables were focused on in a questionnaire. First was the number of books at home (Elley, 1994; Sieben & Lechner, 2019) on a 6-point Likert scale (1: 0–10 books; 2: 11–25 books; 3: 26–100 books; 4: 101–200 books; 5: 201–500 books; 6: > 500 books) to measure the so-called *cultural capital* (P_{11}) (M = 3.21, SD = 1.60). Second, we used three items (Cronbach's alpha = 0.76) to measure *mathematical self-efficacy* (P_{12}) (e.g., "Mathematics is one of my strongest subjects") with a 4-point Likert scale from 1 (not at all true) to 4 (exactly true) (M = 2.35, SD = 0.83). Finally, after each task, the students rated their *contextual prior knowledge* (P_{13}) (e.g., "Did you know anything about varroa mites before?") on a 4-point Likert scale from 1: ("I have never heard of it") to 4 ("I already knew a lot about the topic").

3.3. Sample and data collection

A total of 535 students from two comprehensive schools, S_1 (n = 343) and S_2 (n = 192) participated in this study. S_1 has a rural and S_2 has an urban catchment area of pupils. The students were in grade 9 (n = 266) or grade 10 (n = 269) at the time of data collection, with a mean age of 15.21 years (SD = 0.91).

Among the participating students, 48.60 % were girls, 48.41 % were boys, 3.00 % did not identify as either gender, 72.78 % had no migration backgrounds, 20.85 % had at least one parent born abroad, and 6.37 % were born abroad.

The paper-pencil test was conducted at S_1 on a single school day in 19 classes and S_2 on four consecutive school days in 10 classes. To avoid class effects, different test booklet versions were randomly distributed within the classes. After a short test instruction, each pupil had 80 min to work on their booklet. As there were missing values of up to almost 10 % for individual personal variables, missing values were imputed.

3.4. Data preparation and analysis

To answer the three research questions, various psychometric calculations had to be carried out before the planned analyses. First, as each student could only complete one GTC level item from 15 contexts owing to the large number of tasks and a rotational design had to be used accordingly (Table 1), it was necessary to use probabilistic test procedures from Item Response Theory to map all tasks to a common metric. For this purpose, the difficulty of each item was approximated using one-parametric Rasch scaling and the ConQuest software (version 3.0). Initially, all multiple-choice answers were coded dichotomously (0: incorrect/1: correct) so that an identical rating scale was available for all items. The stochastic independence of the items from one another was ensured using only one item per context per respondent. As part of the Rasch scaling with the IRT software ConQuest, the mean value of the person parameters was set to 0 (constraints = cases) so that the item parameters could be freely estimated. To be able to apply the Rasch model, the tasks also had to fulfill psychometric criteria. Accordingly, items with a point-biserial correlation of <0.2 or with a weighted mean square that was too high or too low (< 0.7 and > 1.25) were excluded from the analyses (Wright, 1994). As all three GTC levels were to be compared for each context, if an item had a too low point-biserial correlation or a misfit, all other items in the context were also excluded. Accordingly, the final Rasch scaling was carried out with a reduced item pool of 22 contexts, each with three items fulfilling the psychometric conditions. As a result of this calibration process, 66 items were included in the analyses to answer the research questions, which had logit values (δ_{1-66}) between −1.65 and 2.67 (Mean 0.61, SD 0.98) as comparable difficulty values. The expected EAP reliability of this scaling can be described as good at 0.77.

Second, to analyze the influence of personal characteristics as possible main and interaction effects, the three variables of cultural capital, mathematical self-efficacy and contextual prior knowledge were used. However, these values were not available for around 10 % of the

test subjects. To be able to perform analyses with a constant number of cases and to avoid possible biases, we approximated the missing values by means of multiple imputation by the mice package (version 3.16.0) in R (version 4.2.1) in the context of 10 imputation runs. The mice package uses an iterative Markov Chain Monte Carlo algorithm for the different runs and the pooling is performed by the robust standard error estimate, so that a complete data set can be generated (Van Buuren, 2018; van Buuren & Groothuis-Oudshoorn, 2011).

Third, the multicollinearity between the mathematical, linguistic and contextual task parameters included in the planned analyses was checked using bivariate correlations with the Benjamini–Hochberg false discovery rate control method (see Table 2).

As correlations below 0.8 were consistently found and the same result was also obtained for the personal variable parameters, no suppression effects were to be expected in joint analyses of these task- and person-related factors (Table 3).

Accordingly, using the lme4 package (version 1.1-31), various generalized linear mixed models (GLMM) could be calculated with these task and person-related variables to answer the three research questions. When analyzing factors influencing the concrete solution of an item as a dependent variable, such models prove necessary owing to the design used in the present project. Their application to a dataset in a long format enables the simultaneous consideration of random effects through common contexts and persons and non-normally distributed responses through dichotomous coding (Berridge et al., 2011).

To answer the first research question, regarding the influence of individual mathematical factors, after descriptive analysis the four mathematical factors were first compared with each other within the framework of four non-nested GLMMs and with a model that included all factors (Section 4.1). While the first three task characteristics represent ordinal measures, the use of dummy coding (three-level indicators with level 1 as reference) was necessary for the nominal variable of cognitive level, as strictly linear effects could not be assumed.

Research question 2 on the influence of other non-mathematical task characteristics and research question 3 on the influence of central personal characteristics on the solution process were then answered (Section 4.2). For this purpose, the mathematical factors (Ma1–4), deliberately not z-standardized for reasons of content interpretation, and the z-standardized linguistic (L5–6), contextual (C7–8) and personal (P9–13) factors were successively integrated into GLMMs as fixed effects. Finally, theoretically justifiable interaction effects of the various factors were considered as interaction terms in the model calculation (Section 4.3).

In all models, the various β coefficients describe the change in the probability of solving a task correctly, and the persons (ID) and tasks (task_ID) were considered as random effects. The analyses were conducted using the *glmer function* of the *lme4* package in *R* (Bates et al., 2015).

Table 3
Means, SDs, and correlations of the personal variables.

Variable	*M*	*SD*	1	2	3	4
1. Intra-mathematical competence (wle)	0.02	1.93				
2. Language competence (wle)	−0.37	1.62	0.00			
3. Cultural capital	3.21	1.60	0.19**	0.02		
4. Mathematical self-efficacy	2.36	0.82	0.36**	0.05	0.14**	
5. Contextual prior knowledge	2.33	0.50	0.14**	0.03	0.18**	0.07

* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

4. Results

4.1. Influence of systematic variation of mathematical factors on empirical difficulty (research question 1)

Table 4 first demonstrates that the intended systematic variation of item difficulty based on the four mathematical task characteristics worked and that three GTC levels could be identified.

Fig. 4 furthermore shows the logit values (δ) of the 66 tasks separately for 22 out of 30 of the remaining real contexts (C01–C30) for each of the three GTC levels. There was a wide dispersion in the range of difficulty spectra within the different contexts; therefore, the mathematical characteristics influenced solution rates differently. Additionally, there were instances when the most difficult task in one context was easier than the easiest task in another context (e.g., C10 and C11).

In addition to the question of the use of mathematical factors for targeted item construction, the main aim was to investigate which mathematical characteristic factor has what influence on the logit value of a realistic task from an item pool with a wide range of difficulty. To this end, we first investigated the extent to which the dichotomously coded individual task responses of the students (dependent variables) could be explained by four different GLMMs, each with a specific mathematical characteristic as an independent variable (Model 1a-1d). In this analysis of the 4576 item responses, the 66 tasks and the 535 individuals were considered as random effects.

Table 5 shows that the mathematical variables of Models 1a, 1c, and 1d initially exhibited significantly negative β coefficients. The probability of solving an item correctly thus decreased significantly with the increase in three of the four task characteristics analyzed, so that these characteristics individually contribute to explaining the variance in task difficulty. However, the joint modeling of the four mathematical variables shows that only the number of procedures and the cognitive level remained as significant influencing variables in a common Model 1. In this context, it also turns out that the cognitive level, coded using dummy indicators, rose sharply from level 1 to 2 ($\beta_{4a} = -0.69$) and this effect weakens only slightly at level 3 ($\beta_{4b} - \beta_{4a} = -1.22 - (-0.69) =$

Table 2
Means, SDs, and correlations of the task characteristics.

Variable	*M*	*SD*	1	2	3	4	5	6	7
1. Number of procedures	2.67	1.71							
2. Difficulty of data selection	0.10	0.23	0.31*						
3. Curricular position	1.36	0.57	0.64**	0.02					
4. Cognitive level	1.71	0.76	0.67**	0.23	0.74**				
5. Number of characters	2109.21	222.45	0.08	−0.03	0.22	0.28*			
6. Linguistic complexity	52.33	4.08	0.02	−0.17	0.03	0.10	0.30*		
7. Interestingness of the context	2.14	0.29	0.01	0.01	0.04	−0.01	0.19	−0.01	
8. Familiarity of the context	2.18	1.24	−0.07	−0.11	−0.09	−0.03	0.22	0.13	0.28*

* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

Table 4
Means and SDs for the Relative Solution Rate and the item difficulty values Logit δ of the Rasch scaling (total and differentiated by GTC level).

	M	*SD*	M_{GTC1}	SD_{GTC1}	M_{GTC2}	SD_{GTC2}	M_{GTC3}	SD_{GTC3}
Solution rate [%]	38.78	17.99	54.14	16.19	37.15	12.12	25.04	12.00
Rasch Logit δ	0.61	0.98	-0.23	0.81	0.68	0.67	1.37	0.73

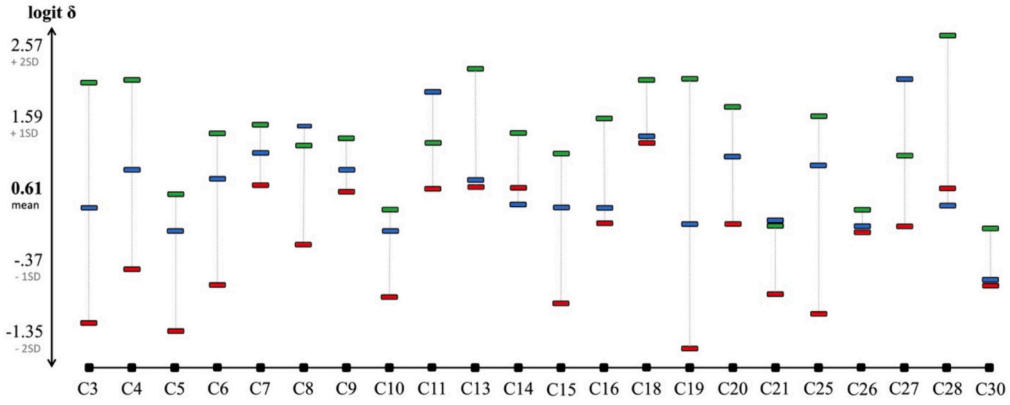


Fig. 4. Item Difficulty values Logit δ of the Rasch scaling of the 66 tasks separated by contexts and GTC level 1 (red), 2 (blue), and 3 (green). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Comparison of the influence of individual mathematical characteristics on item responses.

	Model 1a		Model 1b		Model 1c		Model 1d		Model 1	
	β	SE	β	SE	β	SE	β	SE	β	SE
Fixed effects										
Intercept	**−0.56*–0.56***	0.10	**−0.52***–0.52***	0.10	**0.56*0.56***	0.12	0.06	0.07	0.17	0.26
Mathematical characteristics										
Ma1 Number of procedures	**−0.35***–0.35***	0.05							**−0.21**–0.21****	0.08
Ma2 Difficulty of data selection			−0.59	0.51					0.35	0.41
Ma3 Curricular position					**−0.84***–0.84***	0.18			0.18	0.25
Ma4a_Dummy Cognitive level 2							**−1.02***–1.02***	0.19	**−0.69**–0.69****	0.23
Ma4b_Dummy Cognitive level 3							**−1.56***–1.56***	0.25	**−1.22***–1.22***	0.37
Random effects										
Person variance/ Item variance	0.74 ID/ 0.45 task_ID		0.75 ID/ 0.78 task_ID		0.75 ID/ 0.58 task_ID		0.75 ID/ 0.41 task_ID		0.75 ID/ 0.36 task_ID	
N	4576 responses 535 ID/ 66 task_ID		4576 responses 535 ID/ 66 task_ID		4576 responses 535 ID/ 66 task_ID		4576 responses 535 ID/ 66 task_ID		4576 responses 535 ID/ 66 task_ID	
Model statistics										
Marginal R^2/ Conditional R^2	0.075/ 0.321		0.003/ 0.319		0.046/ 0.320		0.079/ 0.319		0.092/ 0.321	
Deviance	5539.7		5571.2		5553.7		5534.1		5526.8	
AIC	5547.7		5579.2		5561.7		5544.1		5542.8	
BIC	5573.4		5604.9		5587.4		5576.2		5594.3	

* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

−0.53), but still remained significant.

When comparing the five models, there were some differences in terms of the conditional R^2 (Nakagawa et al., 2017), but the difficult choice between the Akaike information criterion (AIC) and Bayesian information criterion (BIC, Burnham & Anderson, 2004) did not initially provide a clear vote. However, the likelihood ratio test (Lewis et al., 2011) ultimately showed that when comparing the nested Models 1 and 1a to 1d, Model 1 had the significantly best model fit ($p < 0.01$) and therefore served as the starting point for answering research questions 2 and 3.

4.2. Additional influence of linguistic, contextual, and personal variables on empirical item difficulty (research question 2)

To answer the second research question, various nested GLMMs were calculated to analyze the influence of linguistic, contextual, and personal characteristics (cf. Table 6).

Based on Model 1, *the number of characters* and *linguistic complexity* of the task texts were integrated within Model 2 as potentially relevant (research question 2a). The results showed that neither ($\beta_{L5} = -0.09$, $p = 0.31$ and $\beta_{L6} = -0.05$, $p = 0.58$) influenced the individual probability of solving an item correctly. Models 1 and 2 showed no differences in fit to the data (χ^2 (2) = 1.76, $p = 0.42$). In this respect, the small variations in the linguistic text features, by deliberately keeping them constant at a medium level (see Section 3.1.1), did not appear to have any influence on the probability of students solving a reality-based task.

In Model 3, the two context characteristics *interestingness* and *familiarity* were also integrated (research question 2b). The output showed a negative correlation, that is, tasks that are generally considered more interesting for students have a lower solution rate ($\beta = -0.24$, $p < 0.001$, odds ratio [OR] = 0.79). In contrast, we found that tasks with contexts with which students should be more familiar had a significantly higher solution rate than tasks rated as rather unfamiliar ($\beta = 0.18$, $p < 0.05$, OR = 1.20). The χ^2 difference test showed that Model 3 fit the data better than Models 1 (χ^2 (4) = 14.03, $p < 0.01$) and 2 (χ^2 (2) = 12.27, $p < 0.01$).

4.3. Personal variables as direct or moderator effect on empirical item difficulty (research question 3)

In Model 4, which had better fit values compared to Model 3 (χ^2 (18) = 209.48, $p < 0.001$), five additional personal characteristics were added to analyze related main and interaction effects (research question 3). In this respect, the interaction effects between individual intra-mathematical competence (P_9) and the mathematical task characteristics ($M_{1–4}$); between individual language competence (P_{10}) and linguistic task characteristics ($L_{5\ \&\ 6}$); and between socioeconomic status (P_{11}), mathematical self-efficacy (P_{12}), contextual prior knowledge, and the context variables ($C_{7\ \&\ 8}$) were analyzed. As theoretically expected, the four variables of intra-mathematical competence ($\beta = 0.32$, $p < 0.05$,

Table 6
GLMMs on the influence of mathematical, language, and personal variables on the probability of solving a task (with moderator effects).

	Model 1		Model 2		Model 3		Model 4	
	β	SE	β	SE	β	SE	β	SE
Fixed effects								
Intercept	0.17	0.26	0.17	0.26	0.02	0.25	0.00	0.25
Mathematical characteristics								
Ma1 Number of procedures	**−0.21*****	0.08	**−0.22*****	0.08	**−0.20*****	0.07	**−0.21*****	0.07
Ma2 Difficulty of data selection	0.35	0.41	0.35	0.41	0.33	0.39	0.29	0.39
Ma3 Curricular position	0.18	0.25	0.18	0.25	0.29	0.23	0.31	0.24
Ma4a_Dummy Cognitive level 2	**−0.69****	0.23	**−0.69****	0.23	**−0.68****	0.22	**−0.72****	0.22
Ma4b_Dummy Cognitive level 3	**−1.22****	0.37	**−1.22****	0.37	**−1.27*****	0.35	**−1.30*****	0.35
Language variables								
L5 Number of characters			−0.09	0.09	−0.09	0.09	−0.07	0.09
L6 Linguistic complexity			−0.05	0.08	−0.08	0.08	−0.06	0.08
Context variables								
C7 Interestingness of the context					**−0.24****	0.07	**−0.26*****	0.08
C8 Familiarity of the context					**0.18***	0.08	0.16	0.09
Personal variables								
P9 Intra-mathematical competence							**0.32***	0.13
P10 Language competence							**0.37*****	0.05
P11 Cultural capital							0.06	0.04
P12 Mathematical self-efficacy							**0.15****	0.05
P13 Contextual prior knowledge							**0.10***	0.04
Moderator effects								
Ma1 × P9 Number of procedures × Intra-mathematical competence							−0.01	0.04
Ma2 × P9 Difficulty of data selection × Intra-mathematical competence							0.03	0.23
Ma3 × P9 Curricular position × Intra-mathematical competence							0.07	0.10
Ma4a × P9 Cognitive level 2 × Intra-mathematical competence							−0.02	0.18
L5 × P10 Number of characters × Language competenceL5 × P10 Number of characters × Language competence							−0.06	0.04
L6 × P10 Linguistic complexity × Language competence							−0.03	0.04
C7 × P11 Interestingness of the context × Cultural capital							0.02	0.04
C8 × P11 Familiarity of the context × Cultural capital							0.00	0.04
C7 × P12 Interestingness of the context × Mathematical self-efficacy							0.05	0.04
C8 × P12 Familiarity of the context × Mathematical self-efficacy							**−0.08***	0.04
C7 × P13 Interestingness of the context × Contextual prior knowledge							0.04	0.04
C8 × P13 Familiarity of the context × Contextual prior knowledge							0.04	0.04
Random effects								
Person variance/Item variance	0.75ID/ 0.36task_ID		0.75ID/ 0.35task_ID		0.75ID/ 0.27task_ID		0.36ID/ 0.27task_ID	
N	4576 responses 535 ID/ 66 task_ID		4576 responses 535 ID/ 66 task_ID		4576 responses 535 ID/ 66 task_ID		4576 responses 535 ID/ 66 task_ID	
Model statistics								
Marginal R^2/Conditional R^2	0.092/ 0.321		0.093/ 0.320		0.108/ 0.320		0.202/ 0.331	
Deviance	5526.8		5525.1		5513.9		5310.0	
AIC	5542.8		5545.1		5537.8		5363.0	
BIC	5594.3		5609.4		5614.0		5562.2	

Note. For the analyses, all L-, C-, and P-variables were z-standardized.
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

OR = 1.37), language competence (β = 0. 37, $p < 0.001$, OR = 1.44), mathematical self-efficacy (β = 0.15, $p < 0.01$, OR = 1.16) and contextual prior knowledge (β = 0.10, p < 0.05, OR = 1.11) significantly contributed to the explained variance. All were significantly positively related to the probability of solving a task. Only the cultural capital (β = 0.11, $p < 0.18$, OR = 1.06) variable did not contribute significantly to the variance explanation, contrary to theoretical assumptions.

Regarding possible interaction effects, task characteristics were uninfluenced by intra-mathematical competence in generating difficulty. Additionally, no correlation was found between the linguistic demands of the tasks and the participants' linguistic competencies; this means that even for weaker students there were no additional linguistic challenges in solving reality-based tasks. Furthermore, no interaction effects were found for socioeconomic background or contextual prior knowledge. Besides, a significant moderator effect emerged regarding familiarity of the context and mathematical self-efficacy. Thus, the negative effect that a task context that is generally rather unfamiliar to students has on the individual probability of solving the task is significantly reduced by mathematical self-efficacy ($\beta_{\text{familiarity of the context} \times \text{mathematical self-efficacy}} = -0.08$, $p < 0.05$, OR = 0.92). Fig. 5 illustrates this effect by showing the solution rates as a function of mathematical self-efficacy for the three context familiarity groups: mean, mean + 1SD, and mean − 1SD.

5. Discussion

We investigated which difficulty-influencing features characterize test items that meet competence-oriented educational requirements. Thus, 90 mathematical items were formulated, with 30 real-world contexts presented in lengthy texts. All were tested by 535 students in a paper-pencil test with a multimatrix test design. In the process, 22 contexts or 66 tasks demonstrated the necessary psychometric qualities and were used for further analyzes. Using GLMM, central task features and moderating variables were identified.

Despite such text-heavy reality-based tasks, a considerable part of the variance in empirical difficulty was explained by mathematical variables describing the mathematical demands. The data demonstrate that while even a single mathematical characteristic could be relatively well suited to describe the difficulty of test items as a global measure, it would be more appropriate to consider a spectrum of mathematical variables. The extent to which this spectrum should be extended beyond the aspects used in this study may be the subject of further research, especially if these results are used for testing and training teachers in diagnostic competence. This is because previous findings indicate that knowledge of differentiated mathematical task features results in teachers significantly diagnosing students' solution processes effectively. This enables more adaptive teaching (Ostermann et al., 2018).

What appeared surprising was the strong role played by the real-world context. While task difficulty can be manipulated using mathematical features, considerable differences exist between task contexts. After all, it seems to show that not only internal mathematical aspects but in particular the context-related formation of the situation model and structuring within the framework of the real model can represent central difficulty-generating characteristics, especially for the processing of reality-based tasks (Leiss et al., 2019).

These findings are supported by the present study because its analyses show that interest in and familiarity with the topic of mathematical tasks significantly influence the solution rate. A less familiar context makes it challenging to form an adequate situation model (Thevenot, 2010), causes a higher cognitive load, and correspondingly reduces the probability of correct task completion. This effect is moderated by

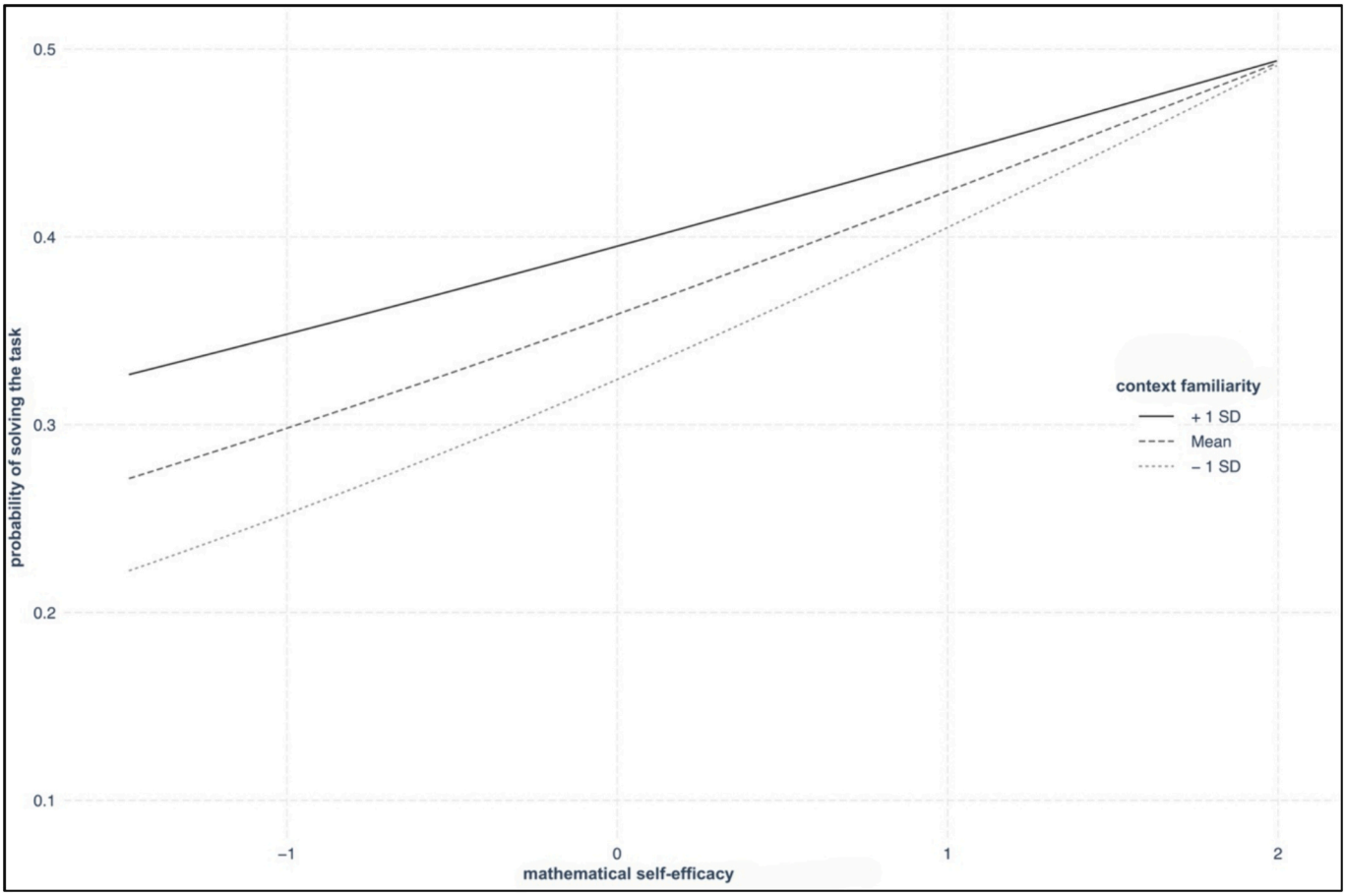


Fig. 5. Moderator effect of context familiarity and mathematical self-efficacy.

mathematical self-efficacy. Students who possess appropriate mathematical self-efficacy also appear to handle more difficult contexts relatively easily, whereas such tasks pose a particular hurdle for students with lower mathematical self-efficacy.

The findings on the interestingness of the task context were contrary to expectations. It is generally assumed that a high level of interest positively affects performance. However, our findings showed that high interestingness was accompanied by a lower solution rate. As the students were only asked about the interestingness of the task after mathematical processing, it can be speculated that those who processed a task at the more challenging GTC level 3 interacted more intensively with the context and therefore evaluated it higher. This also underlines the need for research in mathematics didactics regarding the role of specific contexts in processing reality-based tasks. Particularly, differentiated results are expected concerning the influence of other personal characteristics such as (gender-specific) reading interests (Taylor, 2004) or the ability to use specific comprehension strategies (Boonen et al., 2014).

Conversely, the linguistic features used in the task texts did not influence difficulty. This does not deny the strong connection between linguistic and mathematical competencies (Peng et al., 2020) or the empirical finding that word problems can be particularly challenging for students (Verschaffel et al., 2020). However, the presence of a lifeworld context can also be facilitating (Koedinger & Nathan, 2004). In the processing of tasks with an intermediate reading level among students at a thoroughly advanced reading level, the general linguistic requirements of a task play a subordinate role. Existing findings indicate that the specific interaction of subject-specific linguistic (Vanluydt et al., 2021) and crucial subject requirements (Daroczy et al., 2015) must be examined in a much more differentiated way (Prediger et al., 2019) and extended research methods should be used (Dröse et al., 2021).

Apart from the relationship between context familiarity and mathematical self-efficacy, no further moderating effects occurred between task and personal characteristics; therefore, high- and low-achieving students were equally affected by the mathematical demands of a reality-based task. Linguistically weaker students perform worse in reality-based tasks but do not suffer particularly from linguistic demands, a result that should be further investigated owing to contrary findings (Buono & Jang, 2021). While possessing specific world knowledge helps mathematical engagement with real-world contexts, it cannot compensate for the influence of a lack of interest in or familiarity of a context.

5.1. Limitations

First, our tasks can only partially reflect the real situation in which responsible citizens solve authentic problems using mathematical tools. Thus, both the written contexts and the relevant questions, which should represent a specific variation of mathematical requirements, only represent a processed image of reality. Therefore, students' behavior in real situations is much more complex and influenced by numerous factors. Furthermore, a more systematic use of real-world contexts in the tasks should be aimed at, for example, adapting established models for context characterization (Kerby & Ragan, 2002). In this context, it is also important to systematically vary central factors such as the complexity of forming the situation and reality model, interestingness and familiarity. However, more recent findings on gender-specific influencing factors when dealing with reality-based tasks must be taken into account (Boaler, 1994). Furthermore, only tasks from the area of functional contexts have been analyzed so far; thus, generalization to other areas of mathematics or even other subjects still requires empirical testing.

Second, regarding the variables potentially influencing difficulty, an attempt was made to cover a certain spectrum simultaneously. It would be crucial (in a further experimental step) to systematically vary each variable in isolation.

The third limitation relates to the instruments used. In future studies, it would be better not to use the traditional LIX but to resort to newer procedures, such as the CAREC by Crossley et al. (2019) or, in view of the linguistic specificities, possibly to construct and use more math-related language tests. Additionally, even if it is time efficient, context-related prior knowledge should be recorded as a knowledge test and not as a self-report.

6. Conclusions

This study contributes to research on factors explaining difficulties in mathematics tasks and possibly other subjects, intended for inclusion in a competency-based school curriculum. While many studies have focused on retrospective analyses of individual difficulty factors in existing surveys such as PISA, we used a systematic and theory-based approach in developing mathematical tasks, considering mathematical, linguistic, and contextual features across low, medium, and high difficulty.

In conclusion, the results indicate that (1) within a competence-oriented framework, the focus in both the testing situation and the related teaching situation cannot exclusively be on mathematical aspects; (2) for the field of large-scale assessment as part of educational policy, the range of tasks needs expansion concerning normative goals, and fairness testing should be considered in existing instruments regarding aspects not yet focused on; and (3) further interdisciplinary studies with experimental designs are necessary for this area to better understand and describe the interaction of the various influencing variables.

Declaration of external funding

The data collection and analysis described here was conducted as part of an interdisciplinary project funded by the German Research Foundation (DFG-project no. 417017613). After the end of the project period, the data will be made available to researchers on a national education server for further analysis.

Declaration of ethical approval

All relevant ethical guidelines and principles were carefully considered in the preparation of this scientific article. The conduct of the research, as well as data collection, analysis, and interpretation, was performed in strict adherence to ethical standards to ensure that potential impacts on humans and the environment were minimized. A comprehensive ethical evaluation was conducted prior to the study; this weighed all potential risks and benefits of the research. Any interaction with human participants was voluntary and informed consent was obtained. Participant privacy and confidentiality were always respected, and appropriate measures were taken to maintain anonymity. All study participants provided informed consent. This manuscript has not been published or presented elsewhere in part or in entirety and is not under consideration by another journal. We have read and understood your journal's (ethical) policies, and we believe that neither the manuscript nor the study violates any of these.

CRediT authorship contribution statement

Dominik Leiss: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Timo Ehmke:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Lena Heine:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The author(s) declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Abedi, J. (2011). Language issues in item development. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 391–412). Routledge/Taylor & Francis Group.

Ainley, M., Hidi, S., & Berndorff, D. (2002). Interest, learning, and the psychological processes that mediate their relationship. *Journal of Educational Psychology, 94*(3), 545–561. https://doi.org/10.1037/0022-0663.94.3.545

Albarracín, L., Segura, C., Ferrando, I., & Gorgorió, N. (2022). Supporting mathematical modelling by upscaling real context in a sequence of tasks. *Teaching Mathematics and Its Applications: An International Journal of the IMA, 41*(3), 183–197. https://doi.org/10.1093/teamat/hrab027

Anderson, J. (1983). Lix and Rix: Variations on a little-known readability index. *Journal of Reading, 26*(6), 490–496.

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Ärlebäck, J. B., & Bergsten, C. (2013). On the use of realistic Fermi problems in introducing mathematical modelling in upper secondary mathematics. In *Modeling students' mathematical modeling competencies* (pp. 597–609). Springer. https://doi.org/10.1007/978-94-007-6271-8_52.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Berkowitz, M., & Stern, E. (2018). Which cognitive abilities make the difference? Predicting academic achievements in advanced STEM studies. *Journal of Intelligence, 6*(4), 48. https://doi.org/10.3390/jintelligence6040048

Berridge, D., Crouchley, R., & ProQuest.. (2011). *Multivariate generalized linear mixed models using R*. CRC Press.

Björnsson, C.-H. (1968). *Läsbarhet: Lesbarkeit durch Lix* (Aus dem Schwedischen). Liber.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. In *Handbook I: Cognitive domain*. New York: David McKay.

Blum, W., Drüke-Noe, C., Hartung, R., & Köller, O. (Eds.).. (2016). *Bildungsstandards Mathematik: Konkret*. Cornelsen Verlag Scriptor.

Blum, W., & Leiss, D. (2007). Deal with modelling problems. In C. Haines, P. Galbraith, W. Blum, & S. Khan (Eds.), *Vol. 12. Mathematical modelling: Education, engineering and economics-ICTMA* (pp. 222–231).

Boaler, J. (1994). When do girls prefer football to fashion? An analysis of female underachievement in relation to 'realistic' mathematic contexts. *British Educational Research Journal, 20*(5), 551–564. http://www.jstor.org/stable/1500676.

Boaler, J. (2001). Mathematical modelling and new theories of learning. *Teaching Mathematics and its Applications, 20*, 121–128. https://doi.org/10.1093/teamat/20.3.121

Bong, M., Lee, S. K., & Woo, Y.-K. (2015). The roles of interest and self-efficacy in the decision to pursue mathematics and science. In A. Renninger, S. Hidi, & M. Nieswandt (Eds.), *Interest in mathematics and science learning* (pp. 33–48). American Educational Research Association.

Boonen, A. J., de Koning, B. B., Jolles, J., & Van der Schoot, M. (2016). Word problem solving in contemporary math education: A plea for reading comprehension skills training. *Frontiers in Psychology, 7*, 191. https://doi.org/10.3389/fpsyg.2016.00191

Boonen, A. J. H., van Wesel, F., Jolles, J., & van der Schoot, M. (2014). The role of visual representation type, spatial ability, and reading comprehension in word problem solving: An item-level analysis in elementary school children. *International Journal of Educational Research, 68*, 15–26. https://doi.org/10.1016/j.ijer.2014.08.001

Brown, J. P. (2019). Real-world task context: Meanings and roles. In G. A. Stillman, & J. P. Brown (Eds.), *Lines of inquiry in mathematical modelling research in education* (pp. 53–81). Springer International Publishing. https://doi.org/10.1007/978-3-030-14931-4_4.

Buono, S., & Jang, E. E. (2021). The effect of linguistic factors on assessment of English language learners' mathematical ability: A differential item functioning analysis. *Educational Assessment, 26*(2), 125–144. https://doi.org/10.1080/10627197.2020.1858783

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference. *Sociological Methods & Research, 33*(2), 261–304. https://doi.org/10.1177/0049124104268644

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220. https://doi.org/10.1037/h0026256

Cooper, B., & Dunne, M. (1999). *Assessing children's mathematical knowledge: Social class, sex, and problem-solving*. McGraw-Hill Education (UK).

Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading, 42*(3–4), 541–561. https://doi.org/10.1111/1467-9817.12283

Daroczy, G., Wolska, M., Meurers, W. D., & Nuerk, H.-C. (2015). Word problems: A review of linguistic and numerical factors contributing to their difficulty. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.00348

De Bock, D., Verschaffel, L., Janssens, D., Van Dooren, W., & Claes, K. (2003). Do realistic contexts and graphical representations always have a beneficial impact on students' performance? Negative evidence from a study on modelling non-linear geometry problems. *Learning and Instruction, 13*(4), 441–463. https://doi.org/10.1016/S0959-4752(02)00040-3

De Lange, J. (1996). Using and applying mathematics in education. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education: Part 1* (pp. 49–97). Springer. https://doi.org/10.1007/978-94-009-1465-0_3.

Den Heuvel-Panhuizen, V. (2003). The didactical use of models in realistic mathematics education: An example from a longitudinal trajectory on percentage. *Educational Studies in Mathematics, 54*(1), 9–35. https://doi.org/10.1023/B:EDUC.0000005212.03219.dc

Depaepe, F., De Corte, E., & Verschaffel, L. (2015). Students' non-realistic mathematical modeling as a drawback of teachers' beliefs about and approaches to word problem solving. In B. Pepin, & Roeskin-Winter (Eds.), *From beliefs to dynamic affect systems in mathematics education: Exploring a mosaic of relationships and interactions* (pp. 137–156). Springer. https://doi.org/10.1007/978-3-319-06808-4_7.

Dewolf, T., Van Dooren, W., Ev Cimen, E., & Verschaffel, L. (2014). The impact of illustrations and warnings on solving mathematical word problems realistically. *The Journal of Experimental Education, 82*(1), 103–120. https://doi.org/10.1080/00220973.2012.745468

Ding, H., & Homer, M. (2020). Interpreting mathematics performance in PISA: Taking account of reading performance. *International Journal of Educational Research, 102*, Article 101566. https://doi.org/10.1016/j.ijer.2020.101566

Drijvers, P. (2018). Digital assessment of mathematics: Opportunities, issues and criteria. *Mesure et évaluation en éducation, 41*(1), 41–66. https://doi.org/10.7202/1055896ar

Dröse, J., Prediger, S., Neugebauer, P., Delucchi Danhier, R., & Mertins, B. (2021). Investigating students' processes of noticing and interpreting syntactic language features in word problem solving through eye-tracking. *International Electronic Journal of Mathematics Education, 16*(1). https://doi.org/10.29333/iejme/9674

Drüke-Noe, C., & Kühn, S. (2017). Cognitive demand of mathematics tasks set in European statewide exit exams-are some competences more demanding than others?. In *[Paper presentation]. 10th congress of the European Society for Research in Mathematics Education, Dublin, Ireland*.

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing, 23*(3), 290–325. https://doi.org/10.1191/0265532206lt330oa

El Masri, Y. H., Ferrara, S., Foltz, P. W., & Baird, J.-A. (2017). Predicting item difficulty of science national curriculum tests: the case of key stage 2 assessments. *The Curriculum Journal, 28*(1), 59–82. https://doi.org/10.1080/09585176.2016.1232201

Elia, I., & Philippou, G. (2004). The functions of pictures in problem solving. In *Proceedings of the 28th conference of the International Group for the Psychology of Mathematics Education International group for the psychology of mathematics education, Bergen, Norway*.

Elley, W. B. (1994). *The IEA study of reading literacy: Achievement and instruction in thirty-two school systems*. Pergamon Press.

Embretson, S. E., & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science, 50*(3), 328.

Ferrara, S., Svetina, D., Skucha, S., & Davidson, A. H. (2011). Test development with performance standards and achievement growth in mind. *Educational Measurement: Issues and Practice, 30*(4), 3–15. https://doi.org/10.1111/j.1745-3992.2011.00218.x

Fischer, L., Rohm, T., Carstensen, C. H., & Gnambs, T. (2021). Linking of Rasch-scaled tests: Consequences of limited item pools and model misfit. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.633896

Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M., … Fletcher, J. M. (2006). The cognitive correlates of third-grade skill in arithmetic, algorithmic computation, and arithmetic word problems. *Journal of Educational Psychology, 98*(1), 29–43. https://doi.org/10.1037/0022-0663.98.1.29

Garrett, L., Huang, L., & Charleton, M. C. (2016). A framework for authenticity in the mathematics and statistics classroom. *The Mathematics Educator, 25*(1).

Giere, R. N. (1999). Using models to represent reality. In L. Magnani, N. J. Nersessian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery* (pp. 41–57). Springer US. https://doi.org/10.1007/978-1-4615-4813-3_3.

Gijsbers, D., de Putter-Smits, L., & Pepin, B. (2020). Changing students' beliefs about the relevance of mathematics in an advanced secondary mathematics class. *International Journal of Mathematical Education in Science and Technology, 51*(1), 87–102. https://doi.org/10.1080/0020739X.2019.1682698

Graham, J., Tisher, R., Ainley, M., & Kennedy, G. (2008). Staying with the text: The contribution of gender, achievement orientations, and interest to students' performance on a literacy task. *Educational Psychology, 28*(7), 757–776. https://doi.org/10.1080/01443410802260988

Gravemeijer, K., Stephan, M., Julie, C., Lin, F.-L., & Ohtani, M. (2017). What mathematics education may prepare students for the society of the future? *International Journal of Science and Mathematics Education, 15*, 105–123. https://doi.org/10.1007/s10763-017-9814-6

Haag, N., Heppt, B., Roppelt, A., & Stanat, P. (2015). Linguistic simplification of mathematics items: Effects for language minority students in Germany. *European Journal of Psychology of Education, 30*, 145–167. https://doi.org/10.1007/s10212-014-0233-6

Halim, L., Mohd Shahali, E. H., Iksan, H., & Z.. (2021). Effect of environmental factors on students' interest in STEM careers: The mediating role of self-efficacy. *Research in Science & Technological Education*, 1–18. https://doi.org/10.1080/02635143.2021.2008341

Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research, 70*(2), 151–179. https://doi.org/10.3102/00346543070002151

Julie, C. (2007). Learners' context references and mathematical literacy. In C. Haines, P. Galbraith, W. Blum, & S. Khan (Eds.), *Mathematical modelling* (pp. 195–202). Woodhead Publishing. https://doi.org/10.1533/9780857099419.4.195.

Katz, I. R., Bennett, R. E., & Berger, A. E. (2000). Effects of response format on difficulty of SAT-mathematics items: It's not the strategy. *Journal of Educational Measurement, 37*(1), 39–57. https://doi.org/10.1111/j.1745-3984.2000.tb01075.x

Kerby, D. S., & Ragan, K. M. (2002). Activity interests and Holland's Riasec system in older adults. *The International Journal of Aging and Human Development, 55*(2), 117–139. https://doi.org/10.2190/w0g9-nbyn-h6wc-ltdn

Kim-O, M. (2011). *Analysis of item difficulty and change in mathematical achievement from 6th to 8th grade's longitudinal data*. Georgia Institute of Technology.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E., & Vollmer, H. (2004). *The development of national educational standards. An expertise*. Bundesministerium für Bildung und Forschung.

Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *Journal of the Learning Sciences, 13*(2), 129–164. https://doi.org/10.1207/s15327809jls1302_1

Komor, I., van Vorst, H., & Sumfleth, E. (2021). Students' difficulties arising from mathematical modelling in physical chemistry. *CHEMKON, 30*(5), 176–185. https://doi.org/10.1002/ckon.202100046

Le, L. T. (2007). Effects of item positions on their difficulty and discrimination: A study in PISA Science data across test language and countries. Retrieved from https://research.acer.edu.au/cgi/viewcontent.cgi?article=1001&context=pisa.

Leiss, D., Plath, J., & Schwippert, K. (2019). Language and mathematics — Key factors influencing the comprehension process in reality-based tasks. *Mathematical Thinking and Learning, 21*(2), 131–153. https://doi.org/10.1080/10986065.2019.1570835

Lewis, F., Butler, A., & Gilbert, L. (2011). A unified approach to model selection using the likelihood ratio test. *Methods in Ecology and Evolution, 2*(2), 155–162. https://doi.org/10.1111/j.2041-210X.2010.00063.x

Lubienski, S. T. (2000). Problem solving as a means toward mathematics for all: An exploratory look through a class lens. *Journal for Research in Mathematics Education, 31*(4), 454–482. https://doi.org/10.2307/749653

Marzano, R. J., & Kendall, J. S. (2006). *The new taxonomy of educational objectives*. Corwin Press.

Mullis, I. V., & Martin, M. O. (2017). *TIMSS 2019 assessment frameworks*. ERIC. https://timssandpirls.bc.edu/timss2019/frameworks/.

Mullis, I. V., Martin, M. O., & von Davier, M. (2021). TIMSS 2023 assessment frameworks. https://timssandpirls.bc.edu/timss2023/frameworks/index.html.

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface, 14*(134). https://doi.org/10.1098/rsif.2017.0213

National Assessment Governing Board. (2022). *Mathematics assessement framework for the 2022 and 2024 national assessment of educational progress*. U.S. Department of Education. Retrieved from https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/mathematics/2022-24-nagb-math-framework-508.pdf.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. National Council of Teachers of Mathematics. https://www.nctm.org/Standards-and-Positions/Principles-and-Standards.

Neri, N. C., & Retelsdorf, J. (2022). The role of linguistic features in science and math comprehension and performance: A systematic review and desiderata for future research. *Educational Research Review, 100460*. https://doi.org/10.1016/j.edurev.2022.100460

Niss, M. (2003). Mathematical competencies and the learning of mathematics: The Danish KOM project. In *3rd Mediterranean conference on mathematical education*.

Niss, M., Blum, W., & Galbraith, P. (2007). Introduction. In W. Blum, P. Galbraith, W. Henn, & M. Niss (Eds.), *Vol. 10. Modelling and applications in mathematics education: The 14th ICMI study* (pp. 3–32). Springer US.

Nurjanah, A., & Retnowati, Endah (2018). Analyzing the extraneous cognitive load of a 7 th grader mathematics textbook. *Journal of Physics: Conference Series*, 1097. 012131. 10.1088/1742-6596/1097/1/012131.

OECD. (2012). PISA 2012 — Released mathemaitcs items. Retrieved 230127 from https://www.oecd.org/pisa/test/PISA%202012%20items%20for%20release_ENGLISH.pdf.

OECD. (2019). *PISA 2018 - Assessment and analytical framework*. https://doi.org/10.1787/b25efab8-en

Österholm, M., & Bergqvist, E. (2012). What mathematical task properties can cause an unnecessary demand of reading ability?. In *Norma 11, The Sixth Nordic conference on mathematics education, Reykjavík, Iceland*.

Ostermann, A., Leuders, T., & Nückles, M. (2018). Improving the judgment of task difficulties: Prospective teachers' diagnostic competence in the area of functions and graphs. *Journal of Mathematics Teacher Education, 21*(6), 579–605. https://doi.org/10.1007/s10857-017-9369-z

Pajares, F., & Graham, L. (1999). Self-efficacy, motivation constructs, and mathematics performance of entering middle school students. *Contemporary Educational Psychology, 24*(2), 124–139. https://doi.org/10.1006/ceps.1998.0991

Palm, T. (2008). Impact of authenticity on sense making in word problem solving. *Educational Studies in Mathematics, 67*(1), 37–58. https://doi.org/10.1007/s10649-007-9083-3

Peng, P., Lin, X., Ünal, Z. E., Lee, K., Namkung, J., Chow, J., & Sales, A. (2020). Examining the mutual relations between language and mathematics: A meta-analysis. *Psychological Bulletin, 146*(7), 595. https://doi.org/10.1037/bul0000231

Perrenet, J., & Zwaneveld, B. (2012). The many faces of the mathematical modeling cycle. *Journal of Mathematical Modelling and Application, 1*(6), 3–21.

Piel, S., & Schuchart, C. (2014). Social origin and success in answering mathematical word problems: The role of everyday knowledge. *International Journal of Educational Research, 66*, 22–34. https://doi.org/10.1016/j.ijer.2014.02.003

Pollitt, A., Ahmed, A., & Crisp, V. (2007). The demands of examination syllabuses and question papers. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 166–206). QCA.

Pongsakdi, N., Kajamies, A., Veermans, K., Lertola, K., Vauras, M., & Lehtinen, E. (2020). What makes mathematical word problem solving challenging? Exploring the roles of word problem characteristics, text comprehension, and arithmetic skills. *ZDM, 52*(1), 33–44. https://doi.org/10.1007/s11858-019-01118-9

Prediger, S., Erath, K., & Opitz, E. M. (2019). The language dimension of mathematical difficulties. In A. Fritz, V. G. Haase, & P. Räsänen (Eds.), *International handbook of mathematical learning difficulties: From the laboratory to the classroom* (pp. 437–455). Springer International Publishing. https://doi.org/10.1007/978-3-319-97148-3_27.

Prediger, S., Wilhelm, N., Büchter, A., Gürsoy, E., & Benholz, C. (2018). Language proficiency and mathematics achievement. *Journal für Mathematik-Didaktik, 39*(1), 1–26. https://doi.org/10.1007/s13138-018-0126-3

Pulkkinen, J., Eklund, K., Koponen, T., Heikkilä, R., Georgiou, G., Salminen, J., van Daal, V., & Aro, M. (2022). Cognitive skills, self-beliefs and task interest in children with low reading and/or arithmetic fluency. *Learning and Individual Differences, 97*, Article 102160. https://doi.org/10.1016/j.lindif.2022.102160

Raghubar, K. P., Barnes, M. A., & Hecht, S. A. (2010). Working memory and mathematics: A review of developmental, individual difference, and cognitive approaches. *Learning and Individual Differences, 20*(2), 110–122. https://doi.org/10.1016/j.lindif.2009.10.005

Renninger, K. A., Ewen, L., & Lasher, A. K. (2002). Individual interest as context in expository text and mathematical word problems. *Learning and Instruction, 12*(4), 467–490. https://doi.org/10.1016/S0959-4752(01)00012-3

Russell, M., O'Dwyer, L. M., & Miranda, H. (2009). Diagnosing students' misconceptions in algebra: Results from an experimental pilot study. *Behavior Research Methods, 41*(2), 414–424. https://doi.org/10.3758/BRM.41.2.414

Scheidemann, B., Gasteiger, H., & Puca, R. M. (2022). Effects of affiliation-, achievement-, and power-related topics in mathematical word problems on students' performance, task-related values, and expectancies. *PLoS One, 17*(6), Article e0270116. https://doi.org/10.6084/m9.figshare.16200693.v1

Scheja, B. (2017). The changing cognitive requirement of test tasks in mathematics — A longitudinal study of the Polish middle school examinations. *Didactica Mathematicae, 39*, 101–130.

Schukajlow, S., Leiss, D., Pekrun, R., Blum, W., Müller, M., & Messner, R. (2012). Teaching methods for modelling problems and students' task-specific enjoyment, value, interest and self-efficacy expectations. *Educational Studies in Mathematics, 79*, 215–237. https://doi.org/10.1007/s10649-011-9341-2

Schunk, D. H., & Pajares, F. (2009). Self-efficacy theory. In K. R. Wenzel, & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 35–53). Routledge/Taylor & Francis Group.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*(2), 105–126. https://doi.org/10.1207/s15326977ea1102_2

Sieben, S., & Lechner, C. M. (2019). Measuring cultural capital through the number of books in the household. *Measurement Instruments for the Social Sciences, 1*(1), 1–6. https://doi.org/10.1186/s42409-018-0006-0

Smith, C., & Morgan, C. (2016). Curricular orientations to real-world contexts in mathematics. *The Curriculum Journal, 27*(1), 24–45. https://doi.org/10.1080/09585176.2016.1139498

Stacey, K. (2015). The real world and the mathematical world. In K. Stacey, & R. Turner (Eds.), *Assessing mathematical literacy* (pp. 57–84). Springer.

Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal, 33*(2), 455–488. https://doi.org/10.3102/00028312033002455

Stillman, G. (1998). The emperor's new clothes? Teaching and assessment of mathematical applications at the senior level. In P. Galbraith, W. Blum, G. Booker, & D. Huntley (Eds.), *Mathematical modelling: Teaching and assessment in a technology-rich world* (pp. 243–253). Horwood.

Stillman, G. (2000). Impact of prior knowledge of task context on approaches to applications tasks. *The Journal of Mathematical Behavior, 19*(3), 333–361. https://doi.org/10.1016/S0732-3123(00)00049-3

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and Germane cognitive load. *Educational Psychology Review, 22*(2), 123–138. https://doi.org/10.1007/s10648-010-9128-5

Talsma, K., Schüz, B., Schwarzer, R., & Norris, K. (2018). I believe, therefore I achieve (and vice versa): A meta-analytic cross-lagged panel analysis of self-efficacy and academic performance. *Learning and Individual Differences, 61*, 136–150. https://doi.org/10.1016/j.lindif.2017.11.015

Taylor, D. L. (2004). "Not just boring stories": Reconsidering the gender gap for boys. *Journal of Adolescent & Adult Literacy, 48*(4), 290–298. https://doi.org/10.1598/JAAL.48.4.2

Thevenot, C. (2010). Arithmetic word problem solving: Evidence for the construction of a mental model. *Acta Psychologica, 133*(1), 90–95. https://doi.org/10.1016/j.actpsy.2009.10.004

Thevenot, C. (2017). Arithmetic word problem solving: The role of prior knowledge. In D. C. Geary, D. B. Berch, R. J. Ochsendorf, & K. M. Koepke (Eds.), *Acquisition of complex arithmetic skills and higher-order mathematics concepts* (pp. 47–66). Academic Press. https://doi.org/10.1016/B978-0-12-805086-6.00003-5.

Tout, D., & Spithill, J. (2015). The challenges and complexities of writing items to test mathematical literacy. In K. Stacey, & R. Turner (Eds.), *Assessing mathematical literacy* (pp. 145–171). Springer.

Turner, R., Blum, W., & Niss, M. (2015). Using competencies to explain mathematical item demand: A work in progress. In K. Stacey, & R. Turner (Eds.), *Assessing mathematical literacy* (pp. 85–115). Springer.

Van Buuren, S. (2018). *Flexible imputation of missing data.* CRC Press.

Vanluydt, E., Supply, A.-S., Verschaffel, L., & Van Dooren, W. (2021). The importance of specific mathematical language for early proportional reasoning. *Early Childhood Research Quarterly, 55*, 193–200. https://doi.org/10.1016/j.ecresq.2020.12.003

Verschaffel, L., Greer, B., & De Corte, E. (2000). *Making sense of word problems*. Swets & Zeitlinger.

Verschaffel, L., Schukajlow, S., Star, J., & Van Dooren, W. (2020). Word problems in mathematics education: a survey. *ZDM, 52*(1), 1–16. https://doi.org/10.1007/s11858-020-01130-4

Vondrová, N., Novotná, J., & Havlíčková, R. (2019). The influence of situational information on pupils' achievement in additive word problems with several states and transformations. *ZDM, 51*, 183–197.

Vos, P. (2013). Assessment of modelling in mathematics examination papers: Ready-made models and reproductive mathematising. In G. A. Stillman, G. Kaiser, W. Blum, & J. P. Brown (Eds.), *Teaching mathematical modelling: Connecting to research and practice* (pp. 479–488). Netherlands: Springer. https://doi.org/10.1007/978-94-007-6540-5_41.

Vos, P. (2018). "How real people really need mathematics in the real world"—Authenticity in mathematics education. *Education Sciences, 8*(4), 195. https://doi.org/10.3390/educsci8040195

Vos, P. (2020). Task contexts in Dutch mathematics education. In M. Van den Heuvel-Panhuizen (Ed.), *National reflections on the Netherlands didactics of mathematics: Teaching and learning in the context of realistic mathematics education* (pp. 31–53). Springer International Publishing. https://doi.org/10.1007/978-3-030-33824-4_3.

Voyer, D. (2011). Performance in mathematical problem solving as a function of comprehension and arithmetic skills. *International Journal of Science and Mathematics Education, 9*(5), 1073–1092. https://doi.org/10.1007/s10763-010-9239-y

Walkington, C., Clinton, V., Ritter, S. N., & Nathan, M. J. (2015). How readability and topic incidence relate to performance on mathematics story problems in computer-based curricula. *Journal of Educational Psychology, 107*(4), 1051. https://doi.org/10.1037/edu0000036

Walkington, C., Clinton, V., & Sparks, A. (2019). The effect of language modification of mathematics story problems on problem-solving in online homework. *Instructional Science, 47*, 499–529. https://doi.org/10.1007/s11251-019-09481-6

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen, & L. H. Salganik (Eds.), *Defining and selecting key competencies* (pp. 45–65). Hogrefe & Huber Publishers.

Werning, R., Löser, J. M., & Urban, M. (2008). Cultural and social diversity: An analysis of minority groups in German schools. *The Journal of Special Education, 42*(1), 47–54. https://doi.org/10.1177/0022466907313609

Williams, T., & Williams, K. (2010). Self-efficacy and performance in mathematics: Reciprocal determinism in 33 nations. *Journal of Educational Psychology, 102*(2), 453. https://doi.org/10.1037/a0017271

Wright, B. D. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370–371.

Zheng, R. Z., & Gardner, M. K. (2019). *Memory in education.* Routledge.