DE GRUYTER MOUTON EuJAL 2025; 13(1): 227–251

Lena Heine*, Timo Ehmke and Dominik Leiss

Academic language features in mathematical modelling tasks raise difficulty in reading comprehension for secondary students

https://doi.org/10.1515/eujal-2023-0018

Abstract: School-related reading requirements can be challenging and are entwined with the general ability to perform academically in subjects such as mathematics. However, it is empirically still unclear in how far linguistic requirements affect regular secondary school students ability to interact with subject content. The project reported here investigated the effects of academic German language characteristics on text comprehension difficulty of regular students in a German comprehensive school ("Gesamtschule", 10 % heritage language users) in 25 text-heavy modelling tasks on mathematical functions. For each text, three versions with different amounts of academic lexical and syntactic features were constructed, while the content was not altered. N = 407 regular secondary school students (grades 7–10) solved text comprehension items specifically relevant to the mathematical task solution. General measures of language and mathematical proficiency and cultural capital were elicited. The main findings were that a), considerable numbers of students did in fact experience text comprehension difficulties of mathematically relevant information in mathematical modeling tasks; b), a systematic increase of academic language features in the task texts increased comprehension difficulty even when students characteristics were controlled for; c) higher grade level, mathematical and general language competence, and male gender, predicted strongly the ability to solve the comprehension items correctly; and d) number of books at home and language use at home did not moderate text comprehension after all other variables were controlled.

Keywords: Academic language proficiency, Academic German, language in mathematics, secondary school, assessing reading comprehension

***Corresponding author: Prof. Lena Heine,** Ruhr University Bochum, German Studies, Universitätsstr. 159, 44801 Bochum, Germany, E-Mail: lena.heine@rub.de. https://orcid.org/0000-0003-3569-990X
Prof. Timo Ehmke, Leuphana Universität Lüneburg, Institute of Educational Science, Lüneburg, Niedersachsen, Germany, E-Mail: timo.ehmke@leuphana.de. https://orcid.org/0000-0001-9572-2185
Prof. Dominik Leiss, Leuphana Universität Lüneburg, Institut für Mathematik und ihre Didaktik, Lüneburg, Niedersachsen, Germany, E-Mail: dominik.leiss@leuphana.de. https://orcid.org/0000-0002-7518-820X

Zusammenfassung: Schulische Leseanforderungen stellen eine Herausforderung dar und sind mit der allgemeinen Fähigkeit verbunden, in Sachfächern wie Mathematik akademische Leistungen zu erbringen. Bisher ist empirisch noch unklar, inwieweit sich sprachliche Anforderungen bei Regelschülerinnen und -schülern in der Sekundarstufe auf den Umgang mit fachlichen Inhalten auswirken. Das hier vorgestellte Projekt untersuchte die Auswirkungen von bildungssprachlichen Merkmalen des Deutschen auf das Textverständnis von Regelschülerinnen und -schülern einer deutschen Gesamtschule (10 % verwenden zu Hause (auch) eine Herkunftssprache) in 25 textlastigen Modellierungsaufgaben zu mathematischen Funktionen. Für jeden Text wurden drei Versionen mit unterschiedlich vielen Elementen typisch bildungssprachlicher lexikalischer und syntaktischer Merkmale erstellt. Der Inhalt wurde nicht verändert. N = 407 Schülerinnen und Schüler der Sekundarstufe (Klassenstufen 7–10) lösten dazu Textverständnisaufgaben, die sich speziell auf die Lösung der mathematischen Aufgaben bezogen. Es wurden allgemeine Maße der sprachlichen und mathematischen Kompetenz sowie des kulturellen Kapitals erhoben. Die Hauptergebnisse waren folgende a) Eine beträchtliche Anzahl von Schülern und Schülerinnen hatte Schwierigkeiten, mathematisch relevante Informationen in mathematischen Modellierungsaufgaben zu verstehen; b) ein systematischer Anstieg von bildungssprachlichen Merkmalen in den Aufgabentexten erhöhte die Verständnisschwierigkeit, selbst wenn die Schülermerkmale kontrolliert wurden; c) höhere Klassenstufe, mathematische und allgemeine Sprachkompetenz sowie das männliche Geschlecht sagten die Fähigkeit, die Verständnisaufgaben richtig zu lösen, in starkem Maße voraus; und d) nach Kontrolle aller anderen Variablen beeinflussten die Anzahl der Bücher sowie der Sprachgebrauch im Elternhaus das Textverständnis nicht.

Schlüsselwörter: Bildungssprachliche Fertigkeiten, Deutsche Bildungssprache, Sprache im Mathematikunterricht, Sekundarstufe, Leseverstehen testen

Resumen: Los requisitos de lectura relacionados con la escuela pueden suponer un reto y están entrelazados con la capacidad general para rendir académicamente en asignaturas como las matemáticas. Todavía no está claro empíricamente hasta qué punto los requisitos lingüísticos de los alumnos de secundaria afectan a su forma de abordar los contenidos de las asignaturas. El proyecto que aquí se presenta investigó los efectos de las características lingüísticas educativas del alemán en la comprensión de textos de alumnos regulares de una escuela general alemana (el 10 % (también) utiliza una lengua de herencia en casa) en 25 tareas de modelización basadas en textos sobre funciones matemáticas. Para cada texto, se construyeron tres versiones con diferentes cantidades de características léxicas y sintácticas académicas, mientras que el contenido no se alteró. N = 407 estudiantes regulares de secundaria (7º-10º curso) resolvieron ítems de comprensión de textos específicamente relevantes para la solución de la tarea matemática. Se obtuvieron medidas generales de competencia lingüística y matemática y de capital cultural. Los principales resultados fueron los siguientes: a) un número considerable de estudiantes experimentó dificultades en la comprensión de textos con información matemáticamente relevante en tareas de modelización matemática; b) un aumento sistemático de las características del lenguaje académico en los textos de la tarea incrementó la dificultad de comprensión incluso cuando se controlaron las características de los estudiantes; c) un nivel de estudios más alto, la competencia matemática y lingüística general y el sexo masculino predijeron fuertemente la capacidad de resolver correctamente los ítems de comprensión; y d) el número de libros en casa y el uso del lenguaje en casa no moderaron la comprensión de textos una vez controladas todas las demás variables.

Palabras clave: Competencia lingüística académica, alemán académico, lenguaje en matemáticas, secundaria, evaluación de la comprensión lectora

1 Introduction

Secondary education has operated under the assumption that students have the necessary linguistic skills to comprehend both general academic and subject-specific texts. A range of studies, however, have revealed a great variance in secondary students' reading proficiency, which has been found to strongly relate to their performance in specific subjects (e.g., OECD 2019; Snow 2016; Zhu 2022). Mathematics education is intricately linked with reading skills since it employs mathematical modelling tasks – problems that utilize real-world scenarios presented in text – for learning and assessment (Leiss et al. 2019; Verschaffel et al. 2020). This highlights the need for students to be not just mathematically proficient, but also to possess strong reading comprehension skills in order to effectively understand and solve these types of problems. Accordingly, mathematical modelling tasks have shown to raise difficulty for students compared to tasks without an extensive textual input (Abedi et al. 2003) and mathematical achievement is associated with reading comprehension (Leiss et al. 2010; Wijaya et al. 2014). It is thus plausible to assume insufficient linguistic skills to contribute to achievement problems. Some linguistic characteristics are known to cause strain on reading comprehension and text processing (overview in Heine et al. 2018) – among them, complex, nested sentence structures, infrequent and technical vocabulary, and ambiguous expressions. A number of studies have investigated whether such linguistic features had an impact on students' task performance in mathematic modelling tasks. However, the results were overall inconclusive (Cruz Neri and Retelsdorf, 2022). Some studies showed linguistically weaker learners to perform less successfully on linguistically more demanding mathematical tasks and to profit from linguistic simplification (Abedi and Lord 2010; Prediger et al. 2013; Martiniello 2009; Jorgensen 2011), while others indicated no or only minor effects on task difficulty (Pöhler et al. 2017; Walkington et al. 2019; Kieffer, Lesaux et al, 2009; Haag et al. 2015; Heppt et al. 2015). Still, assuming higher linguistic demands to cause higher difficulty levels for secondary students seem to have face validity both for teachers and students: Ehmke et al. (under review) present evidence that experienced teachers tend to grossly over-estimate linguistic difficulties of mathematic modelling task texts for upper secondary students. Also, Hackemann et al. (2022) results show that students perceive linguistically more complex texts as more difficult – while in fact not showing any considerable comprehension problems when actually processing them, compared to linguistically simpler versions of the same task texts.

The inconclusive results could mirror the fact that the studies used a range of age groups, different linguistic variation methods and different mathematical tasks, and frequently conflated linguistic and subject-specific dimensions. Thus, it is necessary to systematically control linguistic elements of mathematical task texts independent of their mathematical demands if one wants to gain insights into what is difficult for secondary students.

To further understanding in this matter, we report results from a larger interdisciplinary project, *VAMPS*[1], whose main aim is to investigate the impact of linguistic features in task prompts on different levels of mathematical task difficulty (Leiss et al. 2024). The purpose of the present study is to investigate whether systematic linguistic variations of academic language features in mathematical modelling tasks cause comprehension problems for secondary school students and whether certain groups of students are more affected by academic language features than others. The following research questions will be addressed:

RQ1: Do regular middle and upper secondary students (grade 7–10) experience general language comprehension difficulties that could potentially hinder successful processing of mathematical modelling problems?

RQ2: If comprehension difficulties arise, do they vary systematically with different degrees of academic language features in mathematical modelling task texts?

RQ3: Which student characteristics mediate or moderate possible effects of comprehension difficulties?

1 Funded by the German Research Foundation (DFG), project no. 417017613. The project also comprises a twin project in the subject area of physics, cf. Schwippert, Zilz & Hötteke (2024).

2 Background

2.1 Reading comprehension, linguistic demands, and academic language

Reading comprehension is a complex process. It combines data-driven decoding of phonological, orthographical, lexical, syntactical and textual forms with top-down processing such as previous (world, subject-specific and reading-specific, e.g. genre) knowledge, as well as the situational goal of the reading activity, and, if successful, results in a mental model of the information represented in the text (van Dijk and Kintsch 1983; Verhoeven and Perfetti 2008). Comprehension can be impeded on all these levels – for instance, caused by lacking lexical or morphosyntactic knowledge, strain on working memory, problems with focus or the integration of information into a coherent representation, or the lack of background knowledge which the text presupposes. The difficulty of a text can therefore not be defined by textual features alone but only in combination with reader characteristics.

This is closely connected to what can be considered a 'normal' level of proficiency for school students. Being a native speaker does not always coincide with the necessary linguistic skills to succeed in school, because academic registers cannot be acquired in colloquial language use settings alone (Cummins 2000; Schleppegrell 2004; Snow and Uccelli 2012). Academic language is frequently associated with dense information and comparatively rare constructions, such as complex nominal phrases, a high number of sub-clauses per sentence, passives, abstract vocabulary and semantically opaque expressions. These features enhance processing demands even in otherwise proficient language users, because they put stress on working memory or because they are not automatized and increase processing activity for an integration into a mental model (Schleppegrell 2004; Snow and Uccelli 2012; for German: Heine et al. 2018). Importantly, the difference between everyday language and academic language is not categorical, but gradual. For example, while complex sentences and passives are more prevalent in the academic register, they do also occur in everyday language, but much less frequently (Snow and Uccelli 2012). Accordingly, it is possible to create texts that express (near-)identical content with a minimum, or a maximum of linguistic features typical for academic language.

Many studies in the field assume second language learners to be at a disadvantage for the acquisition of academic registers. However, research does not provide any evidence that the use of another language at home or second language status causes lower language achievement in the dominant school language unless it coincides with recent immigration which implies that the learners are still in the process of basic language acquisition (e.g., Montrul 2008; De Houwer 2018). Rather, children from low socioeconomic and low educational family backgrounds experience less familiarization with academic language (Domenech and Krah 2014; Morek 2014).

2.2 Reading comprehension for mathematically relevant mental models

With the increasing integration of real-life problems into mathematics education, general language comprehension (Morgan et al. 2014) and reading (Abedi and Lord 2010) have become a focus of research, which initially dealt with relatively straightforward and assessable situations: *"Alice had 9 cookies. She gave some away. Now she has 4. How many cookies did Alice give away?"* (Reusser 1990).

Since the introduction of educational standards, however, modelling tasks with longer texts are appearing increasingly in mathematics lessons (Leiss et al. 2010; see example in Fig 1). They require the mental construction of a more complex situation model which is situated in a quasi-authentic setting.

Varroa Mites

Most bee populations suffer from Varroa mites, which are parasites. The mites damage the larvae of the bees due to their extremely fast multiplication inside the beehive. Even if there are on average only 50 female mites in the hive at the beginning of spring, by late summer there are already more than 2,500 mites. The mites have a special reproduction cycle that lasts only 12 days. At the beginning of spring, the first cycle of reproduction begins when the 50 female mites attach themselves to nurse bees. These are special bees that are responsible for feeding the bee larvae in their cells. As soon as a nurse bee feeds a larva in a cell, the mite secretly goes with it into the bee larva's cell. The mite remains there even when the bee with a wax cap closes the cell. The female mite now begins to lay eggs in the cell, from which initially only male mites hatch after about 70 hours. More eggs hatch 30 hours later, all of which are female. Next, 150 hours after hatching, these female mites are sexually mature, whereupon the male mites mate with them in the cell. When the nurse bees open the cell lids two days later, the cleaning bees remove all the male and most female mites. However, a total of 3 mated female mites survive per infested cell, completing the first reproductive cycle. Now, these 3 female mites can be transported by nurse bees to the next cells, where the reproduction cycle starts all over again.

Figure 1: Mathematical modelling task text *Varroa mites* from the VAMPS project

In addition to inner-mathematical skills, students' reading ability has been found to be a central predictor for the successful completion of such tasks (Leiss et al. 2019). If students struggle to decode the modelling task linguistically, their chances of solving the task will decrease, even if they are mathematically competent. What is important here is that in a mathematical modelling task, the situation model needs to be construed from the text so that it contains the relevant elements for the underlying *mathematical* problem – so, in addition to general processes of meaning construction, students need to categorize the text information with regard to the mathematical model and the specific operations required for a solution. Prediger et al.

(2018) found that not only reading comprehension impacted on 10 graders' ability to solve mathematical world tasks but translating the text information into the mathematical procedures that the test items induced. So, comprehending a text and comprehending it for mathematical task solution might be based on a similar basis, but are ultimately different processes: While general text comprehension aims at the overall situation or distinct elements described in the text, the construction of an adequate mathematical situation model requires an understanding of specific data and their relationship to each other. For the task presented in Figure 1, it is not so much the textual process of biological reproduction of a varroa mite that is to be understood (Figure 2), but the specific quantitative aspects of the reproduction process and how these aspects interrelate (Figure 3).

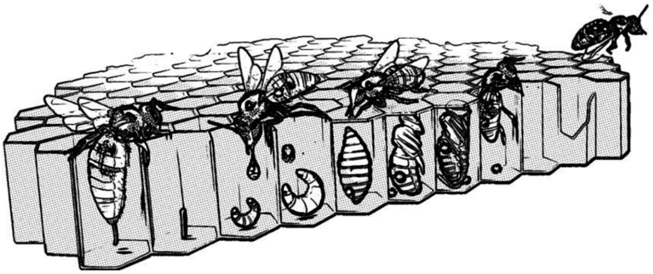

Figure 2: General mental model of the modelling task text *Varroa mites*

Accordingly, the mathematical situation model must be regarded as a specific form of the general one which mathematical task text analysis needs to take into account when identifying potential comprehension difficulties in mathematics tasks.

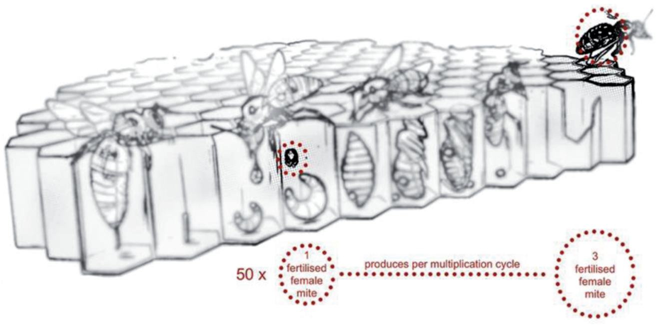

Figure 3: Mathematical mental model of the modelling task text *Varroa mites*

3 Method

3.1 Study design and sample

The study used an experimental design with the within-factor items' academic language level (low=LL1, medium=LL2, high=LL3). The dependent variable is students' reading comprehension of mathematical modelling tasks that are varied in the academic language level. Additionally, several person characteristics as continuous moderator variables were elicited: a) Students' general language proficiency, b) general maths score, and c) number of books at home as an indicator for cultural capital. A convenience sample from a general comprehensive school ("Gesamtschule") in Northern Germany with all academic tracks and a broad range of performance heterogeneity was used. The school is part of a research-practice partnership project and participates regularly in research projects. The sample consisted of a total of N = 424 students from grade 7–10 The sub-samples for each school grade level comprised n=124 students for grade 7, n=125 for grade 8, n=106 for grade 9 and n=69 for grade 10. 52 % of the participants were female, 48 % male. The average age was 14.85 years (*SD*=1.24 years). 10 % of the students speak a different language at home than German. Since the region in which the study was conducted is not characterized by regional varieties, this information must refer to heritage languages spoken due to a migration history of the students' families. 10 % heritage language use in students' homes is a typical distribution for the region. However, the information does not provide any insight into the amount of students' heritage language use, nor their level of competence. Since data were gathered in regular school classes, we can assume that the students understood and spoke German as the language of schooling sufficiently well to follow regular classroom interactions.

Consent to participate was given by the school, parents and students. The students were informed that they would not suffer any disadvantages if they decided not to participate. However, no students declined, presumably because participation was perceived as low-stakes and as an interesting alternative to regular school lessons. The students were tested class-wise in a regular school lesson and those who were present that day participated, so any non-participation was probably due to illness. We therefore estimate the missing values due to non-participation as 'missing completely at random' (MCAR) (Allison 2001) and assume that there is no significant bias of the results. A comparison of the social background indicators of the participating students with indicators from representative school performance studies (OECD 2018) shows that there are no significant differences in the number of books at home. The median of the questionnaire item *books at home* both in our study and in the OECD 2018 German sample is '26-100 books at home' (Mang et al. 2021, p. 29). Therefore, even though we did not opt for a fully representative student sample, we would consider our student sample to be average for the covered age group in Germany.

3.2 Procedure

This study used a rotated paper-pencil test design with 12 different test booklets. Each consisted of the following four parts:

Part 1: Language proficiency test (cloze format) (5 min)

Part 2: Reading comprehension test of mathematical modelling tasks (65 min)

Part 3: Mathematics proficiency test (functional relationships) (10 min)

Part 4: Questionnaire on key demographic personal characteristics (5 min)

The total time of 85 minutes was a usual test duration for the students. While parts 1–4 were identical in all test booklet versions, they differed in terms of the modelling tasks used in part 2. Overall, 25 mathematical modelling tasks were varied each in three levels of the amount of academic language features they contained (LL1, LL2 and LL3, see section 3.2.3). In order to ensure reliability and mitigate text and topic biases, student solved 15 out of the 25 tasks, five tasks each in a LL1, LL2 and LL3 version in a systematic rotation. This ensured that each task version was solved by a large enough sample of students from each year group and each student provided comprehension data from five LL1, LL2 and LL3 texts.

3.2.1 Cloze test as a measure for language proficiency

The students' general German language proficiency was measured with a brief cloze test (5 minutes). The cloze format is well-established as a rather robust reliable and valid measure of general language proficiency (Tremblay 2011). The test used in this study resembles the specific cloze format of the c-test (Grotjahn and Drackert, 2020), but deviates from its canonical structure in that it only used one single text and did not fully adhere to the c-test deletion and scoring measures. The text used in the test consisted of a brief (96 words) nontechnical article extract about health in the workplace, published in the daily newspaper *Hamburger Abendblatt* (2007). There was a balanced proportion of content and function words. In 25 selected words only the first half was given, the second part had to be completed. This ensured an approximation of students' lexical, grammatical, orthographic and textual abilities of both receptive and productive skills. Each gap was coded as a single item (Harsch and Hartig 2015) and scored one point (correct) or 0 points (incorrect or misspelled), yielding a Cronbach's alpha of 0.86. The item scale means were calculated and z-standardized which led to a transformed scale with M=0 and SD=1. The text and deletion technique had in previous studies (Leiss et al. 2018) shown to reliably differentiate samples of regular secondary students in normally distributed data. The cloze test used here can thus be assumed to capture the students' general linguistic abilities in relation to each other.

3.2.2 Mathematical modelling tasks

25 mathematical modelling tasks from the field of mathematical functions were developed by a mathematics expert team. The tasks contained a broad range of quasi-authentic real-world settings, with a broad array of topics, ranging from YouTube influencers' progressive income, rise in fine particle pollution after the New Year's Eve fireworks, to mites in a beehive (Fig. 1). These settings were each presented in a comparably long text prompt which contained all the necessary information students needed to comprehend as a prerequisite for any mathematical task solving process, but also additional information not necessary for the mathematical task. The task designers first wrote rich expository texts and only then designed various related (mathematical) questions. Each text contained several numbers, not all of which relevant for the mathematical model in question. These measures ensured that the mathematically relevant information was not only confined to a short passage of the text and salient, so that reading and understanding the entire text was necessary for the solution of the mathematical task.

3.2.3 Varying the text prompt linguistically

The primary text prompt created by the mathematicians was used by a team of linguists to create three text versions with different linguistic features. First, the text was broken down into single information units that corresponded to the meaning expressed in the original text. Then, a version close to the original text was created (LL2) which adhered to rigid construction principles in order to keep structural elements consistent across all 25 tasks, and which later on served as the starting point for the development of two other versions of the text: One with less features of academic German (LL1), one with more (LL3) than LL2, as illustrated in Table 1.

Table 1: Example mathematical task “Varroa mites” varied in three language level (LL1, LL2 and LL3) text versions

Information units	LL1	LL2	LL3
1	*Die meisten Bienenvölker haben Varroa-Milben.* ‘Most bee populations suffer from Varroa mites.’	*Bei den meisten Bienenvölkern findet man Varroa-Milben, die als Parasiten angesehen werden.* ‘Most bee populations suffer from Varroa mites, which are regarded as parasites.’	*Bei den meisten Bienenvölkern sind die als Parasit anzu*sehende Varroa-Milben zu finden, durch die aufgrund ihrer extrem schnellen Vermehrung im Bienenstock den Larven der Bienen Schaden zugefügt wird. *‘Most bee populations suffer from the Varroa mite, which is to be regarded as a parasite, through whom the bees’ larvae are damaged due to the mites’ extremely quick multiplication inside the beehive.’*
2	*Diese Milben sind Parasiten.* ‘These mites are parasites.’		
3	*Sie schaden den Larven der Bienen.* ‘They damage the bees’ larvae.’	*Die Milben fügen den Larven der Bienen Schaden zu, was man auf ihre extrem schnelle Vermehrung im Bienenstock zurückführen kann.* ‘The mites damage the larvae of the bees due to their extremely fast multiplication inside the beehive.’	
4	*Denn die Milben vermehren sich extrem schnell im Bienenstock.* ‘Because the mites multiply extremely fast inside the beehive.’		
…	…	…	…

Note: English translations are only approximate, not all German syntactic constructions can be formed in analogue to English.

Syntax. The main construction principle of the three text versions was the main clause/sub clause ratio: While LL1 generally used only main clauses, LL2 consistently presented the textual content in a string of main clause + sub clause structures to transport the same informational content. LL1 allowed only subclauses introduced by the highly frequent *weil,* ‘because’, and the complementizer *dass,* ‘that’, when main clause structures would be unidiomatic. LL3, on the other hand, only applied syntactic structures more complex than LL2, and typically exchanged subclauses with participles. Variability in order of information presentation was kept to a minimum across the three text versions; so, one sentence unit in LL2 always integrated two corresponding LL1 sentence units, and LL3 always integrated two corresponding LL2 units (= four LL1 units). Only within each syntactic unit, a different order of content presentation was allowed, which was often necessary due to the more complex syntactic structures used, such as participles or nominalizations, or due to stylistic authenticity.

Verb forms, semantic transparency and verbal/nominal style. Verb forms were a second focus of systematic variation: LL1 used predominantly verb forms instantiated in a single word, while LL2 employed verb forms consisting of several elements; cf. Table 1, information unit 2, which uses *sind* 'are' (in 3rd person plural) in LL1 and the passive structure *angesehen werden* 'are regarded' in LL2. LL3 also employed verb forms consisting of three elements. This construction principle corresponded to the degree of semantic transparency: LL1 used verbs in their prototypical, highly frequent meaning association, such as 'be', 'have', 'damage'; LL2 and LL3, on the other hand, exchanged these verbs with collocational noun-verb combinations which in German can be used as alternatives to simple verbs, often contain particle verbs and are typical and highly frequent features of academic German: In information unit 3, instead of the simple verb *schaden* 'to damage', LL2 uses *Schaden zufügen* 'to provide damage'. Here, *Schaden* 'damage' is instantiating a transition from a verbal (LL1) to a nominal style (LL2 and LL3) and is lexically associated to the particle verb *zufügen* in a collocational way. The example above illustrates that in German, the particle is often split from the verb and put in clause final position, (*fügen ... zu)* which requires the reader, in order to understand the intended meaning, to identify a phrase as a whole and keep all sentential elements in working memory until the clause's end is reached. Care was taken that only verbs were exchanged in this way in which the lexical core could be maintained (*schaden* -> *Schaden zufügen* in the example above) in order to prevent different semantic elements to become prominent in the different text versions. In cases where a more complex verb was necessary and no simple verbs could be found to express a similar meaning, the same lexical core was kept across all three levels. So, information unit 4 in the example above uses the particle verb *vermehren* 'multiply' in LL1 because there is no simple verb which can express the same concept without clouding the intended meaning. This attempts to avoid making LL1 more difficult to comprehend than the more concise LL2. It corresponds to the lexical core in the nominalization *Vermehrung* 'multiplication' in Version LL2 and LL3, which keeps the explicit semantic content stable across all three versions. The frequent use of nominalizations in LL2 and LL3 also have other effects: While the verbal style in LL1 typically realizes concrete agents in the sentences' subject, the nominal style transforms the action itself into the subject. This is a less prototypical function of subjects, which also leads to the necessity to use prepositions and articles in genitive and dative form. Specific care was taken to not alter semantic content; only explicitness of semantic information was varied, e.g., some nouns in LL2 were exchanged with pronouns in LL3.

Deixis. Deictic elements comprise different types of expressions whose meaning can vary in different contexts. Typical deictic elements are personal pronouns or temporal or local expressions such as ‘then’ and ‘there’. In LL1, only such deictica were used for which the necessary relational information was maximally one previous information unit away (Example above, information unit 2 and 3: *Diese Milben ... Sie ...* ‘These mites ... They ...’. In LL2, the corresponding information was never more than one previous sentence (main clause + sub clause) away, while in LL3, the distance could be longer, and could also be cataphoric. This raised the effort of meaning construction across the three versions, because the search for the reference nouns put more pressure on working memory in the more demanding text version.

Impersonal expressions. The final category systematically varied was the use of impersonal expressions which render the actor in a sentence partly or largely invisible, such as *man* ‘one’, as in *man kann sehen, dass ...* ‘one can see that ...’, or passive constructions, which allow to omit the actual agent and are also a typical feature of academic German. While in LL1, such constructions were overall avoided, LL2 and LL3 made use of them.

These principles were applied to the text construction and re-checked several times by a team of linguists until final versions were reached to ensure the typical marker features for each language version were applied, and marker features of higher levels were not included in the lower levels. In cases were semantic or stylistic reasons made it necessary to deviate from the construction principles, the element in question was kept stable across all three text versions in order not to create any variation. Mathematical terminology was considered part of the mathematical test construct and therefore not included into the linguistic variation. Therefore, it was kept identical across the different text versions. The variation led to texts that varied in the range of word count from 140 to 283 but with approximately similarly length across the three text versions (*M*=201 words for LL1, 206 words for LL2, 209 words for LL3).

3.2.4 Creating the mathematical reading comprehension items

In order to determine to which extent linguistic aspects of a mathematics task can account for any potential unsuccessful mathematical processing by the students, specific comprehension test items were developed. These aimed at those comprehension elements that were necessary for processing the mathematics task, rather than a general comprehension of the situation presented in the task text. For this purpose, all modelling tasks were analysed according to the mathematically necessary information they contained. Based on this, nine single-choice items were constructed for each task. For the example task *Varroa mites*, this analysis revealed three necessary information units (number of mites in spring, number of mites needed for a multiplication cycle, number of mites after a multiplication cycle) and two relationships (in spring each mite starts a multiplication cycle and each multiplication cycle does not always add the same number of mites) that students would need to understand for later mathematical processing. Note that in the study presented here, the students did not solve the tasks mathematically but only answered the mathematically relevant comprehension items. For results on a second study with a focus on the mathematical task solution, see Ehmke, Leiss and Heine (under review).

Using only the information in the text, evaluate whether the statement is true or false, or whether it is impossible to judge with only the information given.

		right	wrong	statement not assessable with the info from the text
a)	With each reproductive cycle, exactly 50 mites are added.	☐1	☐2	☐3
b)	One fertilised mite results in three fertilised mites per cycle.	☐1	☐2	☐3
c)	In spring, about 2500 mites live in the hive.	☐1	☐2	☐3
d)	The number of mites triples with each reproductive cycle.	☐1	☐2	☐3
e)	A fertilised mite dies after 3 reproductive cycles.	☐1	☐2	☐3

Figure 4: Comprehension items of the modelling task *Varroa mites*

After a laboratory pilot test had shown that ceiling effects were to be expected in the intended target group, a third answering option, ‘not given in the text’, was included alongside ‘right’ and ‘wrong’, an item format which is known to raise difficulty in reading comprehension tests (Alderson 2000; Castello 2008).

3.2.5 Mathematics proficiency test

Mathematics proficiency was measured by a self-developed, very basic subject test on linear functions. In this short test, all students completed six items in which they had to solve equations (e.g. 5+x=3, x=?). Finally, a sum score was calculated from the tasks solved correctly. The scale reliability was 0.78.

3.2.6 Analysis

For research questions 1 and 2, we calculated the descriptive statistics; further, significant group differences between the LL1 and LL3 task versions were assessed using ANOVA statistics. For all analyses, a significance level of α=.05 was used.

Sato et al.'s (2010) study suggested that applying *Item Response Theory* (IRT) models for estimating the students' achievement scores, compared to using the sum scores or the percentage correct, is more suitable to detect the effects of the academic language levels LL1–LL3. Moreover, for analysing the moderator effects between the personal characteristics (e.g. reading proficiency) and the item attributes (e.g. level of academic features LL1–LL3), explanatory item response models (EIRMs) seemed (Meulders and Xie 2013) adequate and sensitive (De Boeck and Wilson 2013). The EIRM is an approach to cognitive assessment in which explanatory measurements are employed utilising item and person covariates to explain what is being measured, thus adding an explanatory value to the measurements (De Boeck et al. 2016). Thus, for the third research question, we employed the EIRMs that provide a framework for modelling the item responses directly as a function of the item or personal predictors, or a combination of both (De Boeck and Wilson 2013). This makes them especially suitable for the present analysis, as we expected the item responses to depend on an interaction between the items' linguistic characteristics and the students' general language proficiency. In our case, the item responses were binary and we modelled the students' propensity to answer the items correctly, which were the logarithmised odds of a correct answer (e.g., De Boeck and Wilson, 2013). The models were specified as hierarchical Generalised Linear Mixed Models, with item responses nested both within students and items (e.g., De Boeck et al.,2011).

The modification status of the reality-based tasks and the student characteristics (general language proficiency) as well as their interactions were defined as fixed effects. The interaction effects of the general language proficiency (student characteristic) with the linguistic task level (item characteristic) demonstrated differences in the effect of levels of academic language features that are related to the students' general language proficiency (Beretvas et al. 2012; Meulders and Xie 2004). The models were estimated using the lmer function of the lme4 package in R (Bates et al. 2012).

4 Results

Research Question 1 asked whether it was feasible to assume general language comprehension difficulties in regular 7–10 graders that could potentially hinder successful processing of mathematical modelling problems.

On average, the students in our sample solved 56 % of the reading comprehension items correctly. The ability to answer the comprehension items rose significantly with age, and, accordingly, with school grade: While the 13-year-olds solved on average only 51 % correctly (7th grade: 50 %), the percentage rose successively with age, peaking with the 17-year-olds at 70 % on average (10th grade: 71 %), but with a drop to 50 % in the 18-year-olds (a group of only four individuals and an age group not normally found in grade 10). We can thus conclude that there is indeed a considerable number of secondary students who experience difficulties to solve text comprehension items on mathematically important text parts, more so in the younger than in the older secondary students. The systematic increase in correct solution rate from 50 % in grade 7 to 70 % in grade 10 demonstrates that the test instrument we used was overall not too difficult and could capture the range of comprehension abilities in the sample adequately.

To answer Research Question 2, the data were analysed to see whether comprehension difficulties vary systematically across the different text versions LL1, LL2 and LL3. We see a systematic increase of difficulty from texts with fewer features of academic German to those with more such features. This shows that constructions typically associated with academic German can be regarded as difficulty-inducing, and that the model we used for systematic variation delivers consistent results. As Figure 5 shows, it can be concluded that there is indeed an effect of the different amounts of academic language features on text comprehension. Overall, there is a systematic decrease in solution rate from 57 % for LL1, 56 % for LL2 and 53 % for LL3. These differences were statistically significant ($F=14.147$, $p<0.01$). Interestingly, the percentage difference between the three text versions is rather small, and even in the LL1 texts, the students in the sample solved only 57 % of the items correctly.

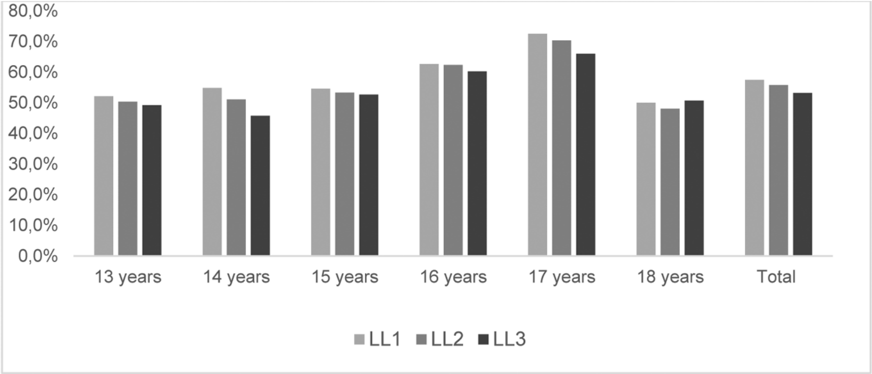

Figure 5: Percent correctly solved reading comprehension items per language level LL1, LL2 and LL3 and age

Research question 3 asked which student characteristics mediate or moderate possible effects of comprehension difficulties.

Plausibly, general language proficiency was a strong positive correlate for the ability to solve the comprehension items: The student quartile with lowest general language proficiency solved 43 %, while the quartile with the highest general language proficiency solved 70 %. Mathematics proficiency also moderated the performance (lowest quartile in mathematics test: 47 % vs. highest quartile 64 %), as well as books at home (0–10 books: 46 % vs. 201–500 books: 58 %).

Student variables originally elicited for sample characterization (gender, language at home) were also analysed. The students who use another language than German at home performed significantly poorer than the students who speak only German (German: 57 % vs. other language: 50 %). A simple comparison between the test scores of boys and girls revealed no statistically significant differences in reading comprehension task performance between males and females in our sample. In order to see whether these results remain stable when all other variables are controlled for and to estimate fixed and random effects of the individual variables, we evaluated four models applying a generalized linear mixed model analysis (GLMM, Table 2). Model 1 analyses the prediction of reading comprehension by students' characteristics only.

Table 2: Prediction of students' reading comprehension by student characteristics and three levels of academic language features (LL1, LL2, LL3) in task texts, including interaction effects

	Model 1		**Model 2**		**Model 3**		**Model 4**	
	β	SE	β	SE	β	SE	β	SE
Fixed effects								
Intercept	0,03	0,10	**0.29*** **	0,10	0,12	0,11	0,17	0,11
Student characteristics								
Gender – female (Reference group)								
Gender – male	**0.26*** **	0,07			**0.26*** **	0,07	**0.19***	0,09
Gender – other	0,21	0,19			0,22	0,19	-0,02	0,24
Language at Home (0 = German, 1 = other than German)	-0,02	0,12			-0,01	0,12	-0,10	0,14
Grade Level	**0.17*** **	0,04			**0.18*** **	0,04	**0.15** **	0,05
Language proficiency	**0.32*** **	0,04			**0.32*** **	0,04	**0.31*** **	0,05
Mathematics proficiency	**0.18*** **	0,04			**0.18*** **	0,04	**0.19*** **	0,05
Books at home	0,04	0,04			0,04	0,04	0,04	0,04

Table 2: (continued)

	Model 1		Model 2		Model 3		Model 4	
	β	SE	β	SE	β	SE	β	SE
Reading items language demands								
LL1 (Reference Group)								
LL2			**-0.08***	0,04	**-0.08***	0,04	-0,09	0,06
LL3			**-0.19*****	0,04	**-0.19*****	0,04	**-0.33*****	0,06
Moderator effects								
Gender – male x LL2							-0,03	0,08
Gender – other x LL2							0,34	0,22
Grade Level x LL2							0,15	0,12
Language at Home x LL2							0,03	0,04
Language proficiency x LL2							0,02	0,05
Mathematics proficiency x LL2							0,01	0,04
Gender – male x LL3							**0.25****	0,08
Gender – other x LL3							0,32	0,22
Grade Level x LL3							0,11	0,12
Language at Home x LL3							0,03	0,05
Language proficiency x LL3							0,02	0,05
Mathematics proficiency x LL3							-0,05	0,04
Random Effects								
τ_{00}	0.36_{id} / 0.20_{task_id}		0.66_{id} / 0.19_{task_id}		0.36_{id} / 0.20_{task_id}		0.36_{id} / 0.20_{task_id}	
N	$20379_{responses}$		$20379_{responses}$		$20379_{responses}$		$20379_{responses}$	
	364_{id} / 25_{task_id}		364_{id} / 25_{task_id}		364_{id} / 25_{task_id}		364_{id} / 25_{task_id}	
Model statistics								
Deviance	25143.1		25303.4		25143.0		25095.7	
AIC	25163.1		25313.4		25119.0		25143.7	

Note. For the analyses, all metric variables were z-standardized. $^{*}p<0.05$ $^{**}p<0.01$ $^{***}p<0.001$

Grade level and general language and mathematics proficiency stay strong and statistically significant predictors of students' reading comprehension in the GLMM analysis: With an increase in these characteristics, the students' reading comprehension increases as well. The GLMM analysis reveals that language use at home loses significance when all other variables are controlled for. An unexpected result is the positive coefficient for male students, which indicates that male students show higher reading comprehension than female students if we control for differences in all other student characteristics.

Model 2 replicates the results from research question 2 above. The coefficients for the levels of academic language in the task texts (Intercept = LL1, LL2, LL3) are statistically significant, which indicates that with increasing number of academic language features (S2, S3) the probability to solve the reading comprehension items decreases.

Model 3 integrates students characteristics and levels of academic language into one single analysis. Comparing the coefficients of model 1 and 2 with the coefficients from model 3 we can observe that none of the coefficients diverts significantly, which indicates that none of the student characteristics (e.g. language proficiency) mediates the effect of linguistic level. Even though the level of academic language features is controlled, gender, grade level and the general proficiencies in language and mathematics remain strong and statistically significant predictors of the students' reading comprehension.

Model 4 tests for all possible moderation effects by including interactions terms between student characteristics and levels of academic language features. The results of the model reveal a positive and statistically significant interaction effect for male students and high levels of academic language features (Gender male x LL3). This indicates that male students are especially successful in solving reading comprehension items compared to female students and under control of all other student characteristics.

5 Discussion and Conclusion

This study set out to investigate whether academic language features in mathematical modelling tasks cause comprehension problems for secondary students, whether such comprehension problems rise with more such language features, and whether some student groups are more affected than others. It was motivated by the lack of conclusive research in the area. The main findings were that for our sample of secondary students in Germany, a), considerable numbers of students did in fact experience text comprehension difficulties of mathematically relevant information in mathematical modeling tasks; b), a systematic increase of academic language features in the task texts increased comprehension difficulty even when students characteristics were controlled for; c) higher grade level, mathematical and general language competence, and male gender, predicted the ability to solve the comprehension items strongly; and d) number of books at home and language use at home did not moderate text comprehension after control of all other variables.

So, for our sample of general secondary students in Germany, a previously under-researched group, we can thus confirm that academic language features do indeed cause an increase in comprehension difficulties. Interestingly, in contrast to expectations from teachers (Ehmke et al. under review) and students (Hackemann et al. 2022) who rated texts manipulated with the same or a similar variation model (Heine et al. 2018) used in the present study, these effects are comparatively small. Since even in the LL1 texts with very few academic language features, the students in the sample solved only 57 % of the items correctly, so the major part of comprehension difficulties could thus not be caused by the academic features we introduced into the different text versions. However, because the increase in difficulty across levels of academic language characteristics remained stable even after controlling for other variables, and because language plays a central role in all school activities, we can hypothesize that cumulative effects may play an important role. Even small comprehension difficulties might accumulate to larger disadvantages over the years. Investigating this empirically is tricky; but longitudinal accounts of learner development could shed light on this and also qualitative study designs could be useful here.

Our data support findings by Abedi and Lord (2010); Martiniello (2009), and Jorgensen (2011) who also found evidence for academic language structures to cause lower mathematical task performance, and Prediger et al. (2013), who confirmed this tendency for German as a target language. Our study is particularly rigid in the construction principles of the mathematical task texts, and because care was taken not to alter the underlying conceptual content, it is possible to provide a valid measure of the effect of linguistic variation without confounding them with content aspects. However, the data do not allow us to distinguish whether the cause of general difficulties in reading comprehension is a lack of basic vocabulary and/or syntactic knowledge, or problems with concentrating and translating word and sentence meanings into a coherent mental model. In addition, for reasons of ecological validity, several academic language structures were manipulated at the same time. Therefore, we cannot determine which effect is caused by which linguistic patterns. Studies are therefore needed to disentangle these aspects with more detailed measures of grammatical competence, working memory, concentration, comprehension effects of individual academic language features, such as comprehension of passives, complex nominal phrases, or other typical academic sentence patterns, especially with respect to the secondary school group.

Our study suggests that academic language features may hinder learning in the subject: Because comprehension tasks were used that targeted mathematical modeling, we can assume that students with comprehension difficulties would also show a poorer mathematical performance. Whether text features do actually lead to lower mathematical performance, however, cannot be answered with the present study, because this would require measurements not only of text comprehension but also of mathematical task solving. Also, because we used mathematical task texts from the same cognitive process domains of mathematical functions, we do not know if the linguistically induced difficulty of task texts is a stable feature or if it interacts with the mathematical domain of the texts. This interrelation will be investigated more systematically in a follow-up study within the *VAMPS* project.

Because home use of a heritage language did not predict poorer comprehension in our data, our study supports theories of heritage language and multilingualism research that do not assume that multilingual students per se are less proficient than monolingual students in the dominant language of society (e.g., Montrul 2008; De Houwer 2018). This is relevant because research in the area of school achievement still tends – which we believe is problematic – to view multilingual students as a problem group with particular obstacles. Our research contributes to making it increasingly clear that multilingualism and heritage language use are not particularly relevant variables in the context of school achievement and should not be viewed as inherently problematic.

A rather surprising finding from our GLMM analysis was that male gender was a strong predictor of high reading comprehension when all other student characteristics were controlled for. This contrasts findings on students' reading performance, where females frequently outperform males (e.g., OECD 2019). Our results could suggest that the mathematical embeddedness of our tasks favors boys over girls and that reading comprehension for mathematical modeling may be a different construct than general reading comprehension. The GLMM analysis also revealed that the number of books in the home, an otherwise very reliable predictor of general language proficiency and school achievement, had no predictive power for comprehension of solving mathematical problems in our sample. This indicates that academic language, contrary to general language competence, might be a general difficulty-inducing feature which affects all secondary students, not only students from lower educational backgrounds. This is in line with the finding that while academic features increase the level of difficulty, even texts with very simple linguistic structures cause problems for many students in comprehension for mathematical task solution, even as late as grade 10. The increase in the level of competence with increasing age and grade level indicates that the general competence to comprehend texts develops considerably in secondary school.

To conclude, our findings provide an indication of the normal range of linguistic diversity in general education classrooms that needs to be recognized and addressed pedagogically without pathologizing this heterogeneity. They confirm once more how instrumental text comprehension ability is in mathematics, and that general students' effortless text comprehension cannot be taken for granted but is developing considerably in secondary school. As the data show that not all students master these skills on their own, systematic academic language development, but also basic text competence training should be an integral part of schooling in all subjects.

References

Abedi, Jamal, Seth Leon & Jim Mirocha. 2003. *Impact of student language background on content-based performance: Analyses of extant data.* University of California, Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Abedi, Jamal & Carol Lord. 2010. The language factor in mathematics tests. *Applied Measurement in Education* 14(3). 219–234. https://doi.org/10.1207/S15324818AME1403_2.

Alderson, J. Charles. 2000. *Assessing reading.* Cambridge: Cambridge University Press.

Allison, Paul D. 2001. *Missing Data.* Sage Publishing. https://doi.org/10.4135/9781412985079

Bates, Douglas, Martin Maechler & Ben Bolker. 2012. lme4: Linear mixed-effects models using S4 classes. R package version 0.999999–0.

Beretvas, S. Natasha, Stephanie W. Cawthon, L. Leland Lockhart & Alyssa D. Kaye. 2012. Assessing impact, DIF, and DFF in accommodated item scores: a comparison of multilevel measurement model parameterizations. *Educational and Psychological Measurement* 72(5). 754–773. https://doi.org/10.1177/0013164412440998.

Castello, Erik. 2008. *Text complexity and reading comprehension tests*. Bern: Peter Lang.

Cruz Neri, Nadine & Jan Retelsdorf. 2022. The role of linguistic features in science and math comprehension and performance: A systematic review and desiderata for future research. *Educational Research Review 36*.100460. https://doi.org/10.1016/j.edurev.2022.100460.

Cummins, Jim. 2000. *Language, power and pedagogy: Bilingual children in the crossfire* (Bilingual Education & Bilingualism). Clevedon: Multilingual Matters. https://doi.org/10.21832/9781853596773.

De Boeck, Paul, Marjan Bakker, Robert J. Zwitser, Michel Nivard, Abe Hofman, Francis Tuerlinckx & Ivailo Partchev. 2011. The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software* 39(12). 1–28. https://doi.org/10.18637/jss.v039.i12.

De Boeck, Paul, & Mark Wilson (eds.). 2013. *Explanatory item response models: A generalized linear and nonlinear approach* (Statistics for Social Science and Public Policy). New York: Springer. https://link.springer.com/book/10.1007/978-1-4757-3990-9 (accessed 21 May 2024).

De Cat, Cécile. 2020. Predicting language proficiency in bilingual children. *Studies in Second Language Acquisition* 42(2). 279–325. https://doi.org/10.1017/S0272263119000597

De Houwer, Annick. 2018. The role of language input environments for language outcomes and language acquisition in young bilingual children. In David Miller, Fatih Bayram, Jason Rothman & Ludovica Serratrice (eds.), *Bilingual cognition and language: The state of the science across its subfields* (Studies in Bilingualism 54), 127–154. Amsterdam: John Benjamins.

Dijk, Teun A. van & Walter Kintsch. 1983. *Strategies of discourse comprehension.* New York: Academic Press.

Domenech, Madeleine & Antje Krah. 2014. What familial aspects matter? Investigating argumentative competences of learners at the beginning of secondary schooling in the light of family-based resources. *Learning, Culture and Social Interaction* 3(2). 77–87. https://doi.org/10.1016/j.lcsi.2013.12.006.

Ehmke, Timo, Dominik Leiss & Lena Heine (under review). The effects of linguistic complexity of reality-based mathematical tasks: discrepancy between teachers' expectations and student's performance.

Hackemann, Timo, Lena Heine & Dietmar Höttecke. 2022. Challenging to read, easy to comprehend? Effects of linguistic demands on secondary students' text comprehension in physics. *International Journal of Science and Mathematics Education.* https://doi.org/10.1007/s10763-022-10306-1

Haag, Nicole, Birgit Heppt, Alexander Roppelt & Petra Stanat. 2015. Linguistic simplification of mathematics items: Effects for language minority students in Germany. *European Journal of Psychology and Education* 30. 145–167. https://doi.org/10.1007/s10212-014-0233-6.

Harsch, Claudia & Johannes Hartig. 2016. Comparing C-tests and Yes/No vocabulary size tests as predictors of receptive language skills. *Language Testing* 33(4). 555–575. https://doi.org/10.1177/0265532 2155946.

Heppt, Birgit, Nicole Haag, Katrin Böhme & Petra Stanat. 2015. The role of academic language features for reading comprehension of language-minority students and students from low-SES families. *Reading Research Quarterly* 50(1). 61–82. https://doi.org/10.1002/rrq.83.

Heine, Lena, Madeleine Domenech, Lisa Otto, Astrid Neumann, Michael Krelle, Dominik Leiss, Dietmar Hötteke, Timo Ehmke & Knut Schwippert. 2018. Modellierung sprachlicher Anforderungen in Testaufgaben verschiedener Unterrichtsfächer: Theoretische und empirische Grundlagen. *Zeitschrift für Angewandte Linguistik*, 69/2, 69–96. https://doi.org/10.1515/zfal-2018-0017

Jorgensen, Robyn. 2011. Language, culture and learning mathematics: A Bourdieuian analysis of indigenous learning. In Claire Wyatt-Smith, John Elkins & Stephanie Gunn (eds.), *Multiple Perspectives on Difficulties in Learning Literacy and Numeracy*, 315–329. Dordrecht: Springer.

Kieffer, Michael J., Nonie K. Lesaux, Mabel Rivera & David J. Francis. 2009. Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research 79*(3), 1168–1201. http://www.jstor.org/stable/40469092 (accessed 21 Mai 2024).

Knabe, Melina L., Emma Lazaroff & Haley A. Vlach. 2022. Academic vocabulary as a measure of school readiness. In Anastasia Lynn Betts & Khanh-Phuong Thai (eds.), *Handbook of research on innovative approaches to early childhood development and school readiness*, 113–141. IGI Global. https://doi.org/10.4018/978-1-7998-8649–5.ch006.

Leiss, Dominik, Stanislav Schukajlow, Werner Blum, Rudolf Messner & Reinhard Pekrun. 2010. The role of the situation model in mathematical modelling: Task analyses, student competencies, and teacher interventions. *Journal für Mathematik-Didaktik 31*(1), 119–141. https://doi.org/10.1007/s13138-010-00 06-y

Leiss, Dominik, Jennifer Plath & Knut Schwippert. 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

Leiss, Dominik, Ehmke, Timo & Heine, Lena. 2024. Reality-based tasks for competency-based education: The need for an integrated analysis of subject-specific, linguistic, and contextual task features. *Learning and Individual Differences*, *114*, 102518. https://doi.org/10.1016/j.lindif.2024.102518.

Mang, Julia; Ludwig Seidl; Anja Schiepe-Tiska; Ana Tupac-Yupanqui; Lisa Ziernwald; Anastasia Doroganova; Mirjam Weis; Jennifer Diedrich; Jörg-Henrik Heine; Elisabeth González Rodríguez & Kristina Reiss. 2021. *PISA 2018 Skalenhandbuch. Dokumentation der Erhebungsinstrumente. Münster, New York : Waxmann.* https://doi.org/10.25656/01:23923

Martiniello, Maria. 2009. Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment* 14(3/4). 160–179. https://doi.org/10.1080/10627190903422906.

Masrai, Ahmed & James Milton. 2018. Measuring the contribution of academic and general vocabulary knowledge to learners' academic achievement. *Journal of English for Academic Purposes* 31. 44–57. https://doi.org/10.1016/j.jeap.2017.12.006.

Meulders, Michel & Yiyu Xie. 2013. Person-by-item predictors. In Paul De Boeck & Mark Wilson (eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (Statistics for Social Science and Public Policy), 213–240. New York: Springer. https://doi.org/10.1007/978-1-4757-3990-9_7.

Montrul, Silvina A. 2008. *Incomplete acquisition in bilingualism: Re-examining the age factor* (Studies in Bilingualism 39). Amsterdam: John Benjamins. https://doi.org/10.1075/sibil.39.

Morek, Miriam. 2014. Constructing social and communicative worlds – The role of peer-interactions in preadolescents' discursive development. *Learning, Culture and Social Interaction* 3(2). 121–133. https://doi.org/10.1016/j.lcsi.2014.02.005.

Morgan, Candia, Tracy Craig, Marcus Schuette & David Wagner. 2014. Language and communication in mathematics education: an overview of research in the field. *ZDM Mathematics Education* 46. 843–853. https://link.springer.com/article/10.1007/s11858-014-0624-9 (accessed 26 June 2023).

OECD (ed.). 2019. *PISA 2018 results: Where all students can succeed: Volume II.* Paris: OECD Publishing. https://doi.org/10.1787/b5fd1b8f-en.

Prediger, Susanne, Nadine Renk, Andreas Büchter, Erkan Gürsoy & Claudia Benholz. 2013. Family background or language disadvantages? Factors for underachievement in high stakes tests. In Anke Lindmeier & Aiso Heinze (eds.), *Proceedings of the 37th conference of the international group for the psychology of mathematics education*, 49–56. Kiel: PME.

Prediger, Susanne, Nadine Wilhelm, Andreas Büchter, Erkan Gürsoy & Claudia Benholz. 2018. Language proficiency and mathematics achievement: Empirical study of language-induced obstacles in a high stakes test, the central exam ZP10. *Journal für Mathematik-Didaktik* 39(2). 1–26. https://doi.org/10.1007/s13138-018-0126-3.

Pöhler, Birte, Ann Cathrice George, Susanne Prediger & Henrike Weinert. 2017. Are word problems really more difficult for students with low language proficiency? Investigating percent items in different formats and types. *International Electronic Journal of Mathematics Education* 12(3). 667–687. https://doi.org/10.29333/iejme/641.

Reusser, Kurt. 1990. Understanding word arithmetic problems. Linguistic and situational factors. Paper presented at the Annual Meeting of the American Education Research Association, Boston, 6–20 April.

Sato, Edynn, Stanley Rabinowitz, Carole Gallagher & Chun-Wei Huang. 2010. *Accommodations for English language learner students: The effect of linguistic modification of math test item sets.* (NCEE 2009–4079). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Schleppegrell, Mary. J. 2004. *The language of schooling: A functional linguistics perspective*. London: Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410610317.

Snow, Marguerite A. 2016. Content-based language teaching and academic language development. In Eli Hinkel (ed.), *Handbook of research in second language teaching and learning: Volume III*, 159–172. New York: Routledge. https://doi.org/10.4324/9781315716893.

Snow, Catherine E. & Paola Uccelli. 2012. The challenge of academic language. In David R. Olson & Nancy Torrance (eds.), *The Cambridge handbook of literacy*, 112–133. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511609664.008.

Schwippert, Knut, Zilz, Kendra & Hötteke, Dietmar (2024). Difficulty-generating features oft ext-based physics tasks. *Journal of Research in Science, Mathematics and Technology Education 7*(2), 1–24. https://doi.org/10.31756/jrsmte.721

Tremblay, Annie. 2011. Proficiency assessment standards in second language acquisition research: "Clozing" the gap. *Studies in Second Language Acquisition* 33(3). 339–372. https://doi.org/10.1017/S0272263111000015.

van Dijk, Teun A. & Walter Kintsch. 1983. *Strategies of discourse comprehension. New York: Academic Press.*
Verhoeven, Ludo & Charles Perfetti. 2008. Advances in text comprehension: model, process and development. *Applied Cognitive Psychology* 22(3). 293–301. https://doi.org/10.1002/acp.1417.
Verschaffel, Lieven, Stanislaw Schukajlow, Jon Star & Wim Van Dooren. 2020. Word problems in mathematics education: a survey. *ZDM Mathematics Education* 52. 1–16. https://doi.org/10.1007/s11858-020-01130-4.
Walkington, Candace, Virginia Clinton & Pooja Shivraj. 2018. How readability factors are differentially associated with performance for students of different backgrounds when solving mathematics word problems. *American Educational Research Journal* 55(2). 362–414. https://doi.org/10.3102/0002831217737028.
Wijaya, Ariyadi, Marja van den Heuvel-Panhuizen, Michiel Doorman & Alexander Robitzsch. 2014. Difficulties in solving context-based PISA mathematics tasks: An analysis of students' errors. *The Mathematics Enthusiast* 11(3). 555–584. https://doi.org/10.54870/1551–3440.1317.
Wilson, Mark & Paul De Boeck. 2013. Descriptive and explanatory item response models. In Paul De Boeck & Mark Wilson (eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (Statistics for Social Science and Public Policy), 43–74. New York: Springer. https://doi.org/10.1007/978-1-4757-3990-9_2.
Zhu, Yuanze. 2022. Reading matters more than mathematics in science learning: an analysis of the relationship between student achievement in reading, mathematics, and science. *International Journal of Science Education* 44(1). 1–17. https://doi.org/10.1080/09500693.2021.2007552.