frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 09 December 2025
DOI 10.3389/feduc.2025.1699117

OPEN ACCESS

EDITED BY
Lutfi Incikabi,
Kastamonu Universitesi Egitim Fakultesi, Türkiye

REVIEWED BY
Ezgi Mor,
Kastamonu University, Türkiye
Oktay Erbay,
Hatay Mustafa Kemal University, Türkiye

*CORRESPONDENCE
Eileen Klotz
✉ Eileen.klotz@leuphana.de

RECEIVED 04 September 2025
ACCEPTED 03 November 2025
PUBLISHED 09 December 2025

CITATION
Klotz E, Ehmke T and Leiss D (2025) Modifying gap placement, topic, and evaluation method: the impact of modified C-Tests and gender on test performance and the relationship between modified C-Tests and mathematical performance. *Front. Educ.* 10:1699117. doi: 10.3389/feduc.2025.1699117

Modifying gap placement, topic, and evaluation method: the impact of modified C-Tests and gender on test performance and the relationship between modified C-Tests and mathematical performance

Eileen Klotz[1]*, Timo Ehmke[1] and Dominik Leiss[2]

[1]Institute of Educational Sciences, Leuphana University, Lüneburg, Germany, [2]Institute of Mathematics and its Didactics, Leuphana University, Lüneburg, Germany

Introduction: C-Tests are employed globally to evaluate general language proficiency, and modifications of the typical C-Test have been proposed. Given the strong connection between language and mathematics, this study, besides investigating the difficulty of C-Test modifications and the impact of the test-taker's gender, examines the relationship between modified C-Tests and mathematical performance across different content areas.
Methods: 190 seventh-graders were surveyed on their mathematical skills and given C-Tests with variations in gap placement (beginning vs. end) and topic (mathematical vs. general). Two evaluation methods were examined (correct/incorrect vs. word-recognition).
Results: GLMM analyses show that (a) the initial gap is the key factor contributing to test difficulty, (b) girls outperform boys, (c) boys benefit more from mathematical text topics, girls from gaps at the end of words, (d) arithmetic shows the strongest, geometry the weakest relationship to C-Test performance, (e) students with higher mathematical skills benefit significantly from mathematical text topics, and (f) students with lower mathematical skills benefit significantly from gaps at the end of words.
Discussion: Findings show that formal characteristics and gender impact C-Test performance and must be kept in mind when determining language proficiency through it. Furthermore, this study extends current research by demonstrating how specific language skills measured through modified C-Tests relate to different mathematical skills and therefore once more advocates the integration of language in mathematics education by presenting the C-Test as a possibility for this.

KEYWORDS

C-Test, gender differences, language and mathematics, language proficiency, language testing, mathematics education

1 Introduction

Since their introduction in 1981, C-Tests have become one of the most extensively studied instruments for measuring general language proficiency (Grotjahn, 2019; Raatz and Klein-Braley, 1982). Based on the reduced redundancy principle (Klein-Braley, 1997), they require test-takers to reconstruct parts of words in short texts and have

thereby been shown to provide a reliable and efficient assessment of general language proficiency in terms of integrative skills that cover all four language skills (e.g., Asano, 2014; Eckes and Grotjahn, 2006). Due to its advantages and potential, the C-Test is commonly employed in placement procedures, language acquisition research, and educational monitoring (e.g., Grotjahn, 1995, 2014). Over time, various modifications of the typical canonical C-Test, in which the second half of every second word is deleted, have been proposed to capture different facets of language proficiency and to adjust test difficulty. These modifications vary in terms of gap placement, deletion principle, or text type and have been shown to systematically affect test difficulty and moreover the type of language skills being assessed (e.g., Höttecke et al., 2017; Sigott and Köberl, 1996; Lee et al., 2019). Furthermore, while some studies point to systematic gender differences in C-Test performance (e.g., Grotjahn et al., 2002; Mashkovskaya, 2013), the evidence remains inconsistent (Tabatabaei and Shakerin, 2013). Despite the ongoing research, systematic investigations that simultaneously consider different C-Test characteristics and test-taker variables such as gender are still missing.

The relevance of such investigations further becomes evident when looking beyond language education. Language proficiency is a decisive factor for success across various subjects, including mathematics. A large body of research highlights the close interplay between language and mathematical proficiency (e.g., Fuchs et al., 2006; Paetsch et al., 2015; Peng et al., 2020; Ufer and Bochnik, 2020). While much research focuses on the importance of language in text-heavy tasks (e.g., Klotz et al., 2025), studies also show that language proficiency can support performance in tasks with low language demands, such as arithmetic (e.g., Berg, 2008; Fuchs et al., 2008), which underpins its relevance for mathematical learning more broadly. Correlations between canonical C-Tests and mathematical performance have already been demonstrated (Brunner et al., 2022). These findings suggest that C-Tests capture language-related skills that are central to mathematical skills in different content areas. However, what remains unclear is how modified C-Tests relate to mathematical skills across different content areas. Because different C-Test modifications measure different language skills (Baur and Spettmann, 2010; Beinborn, 2016; Mashkovskaya, 2013), carefully designed modifications may not only influence test difficulty but also provide more precise insights into which language skills are most relevant for mathematical success. This finding can potentially broaden the applicability of C-Tests, as the close connection between language and mathematics suggests that C-Tests could be integrated into mathematics education to foster mathematical skills.

The present study seeks to address these research gaps. Its aim is twofold. First, it investigates how different characteristics of the C-Test—gap placement, text topic, and evaluation method—influence test performance and whether there are differences between girls and boys. Second, it examines how students' performance in different mathematical content areas is related to their C-Test performance and how this relationship is shaped by specific C-Test characteristics. Accordingly, the study is guided by the following research questions and hypotheses:

RQ1. How do different characteristics of the C-Test affect test performance?

H1. Based on the applied method, it is expected that C-Tests evaluated using the CI-method will show a lower rate of correct solutions. Furthermore, it is assumed that C-Tests with gaps at the beginning of words are more difficult. Regarding the text topic, it is hypothesized that mathematical C-Tests will yield lower solution rates due to the higher demands on technical language.

RQ2. Are there differences between girls and boys?

H2. It is assumed that girls generally perform better in C-Tests due to their stronger literacy skills. However, boys are expected to particularly benefit from the mathematical text topics, as their typically stronger mathematical skills may be reflected in better performance on these tests.

RQ3. How is students' mathematical performance in different content areas related to their C-Test performance, and how is this relationship influenced by the C-Test characteristics gap placement, text topic, and evaluation method?

H3. The strongest relationship between C-Test and mathematical performance is expected to occur in word problems, as these tasks are particularly text-heavy. Owing to the central role of technical language, it is further assumed that C-Tests with mathematical text topics will show a stronger relationship with mathematical performance than C-Tests with general text topics. Moreover, C-Tests evaluated using the CI-method are expected to exhibit a stronger relationship with mathematical performance, as they reflect a higher level of language proficiency. No specific effects are anticipated regarding gap placement.

2 Theoretical background

This section examines the significance and structure of C-Tests in more detail and draws on previous research findings, including C-Test modifications and gender differences. Furthermore, the connection between mathematics and language is highlighted and thereby establishes a link to the significance of this study.

2.1 Origin and structure of C-Tests

The C-Test was first invented in 1981 by Klein-Braley and Raatz (see Raatz and Klein-Braley, 1982) as a global instrument for determining general language proficiency and is one of the most extensively researched language tests since then (e.g., Baur et al., 2006; Grotjahn, 2019; Grotjahn et al., 2002). C-Tests are based on a variation of the cloze principle and therefore share the same theoretical assumption of the concept of reduced redundancy testing (e.g., Klein-Braley, 1997). Here, the deletion of words or part of words is based on the assumption that disturbances can occur during the transmission of information through speech, which is why the redundancy necessary for communication is reduced. The higher the individual's language skills are, the better they can reconstruct the respective text by filling in the gaps correctly (e.g., Grotjahn, 2019). Still, there has been criticism of classical cloze tests in the past, e.g., because of its deletion principle (every 5th to 10th word) and the tests only constisting of one long text,

possibly resulting in test bias. Moreover, the overall reliability of cloze tests is much lower than previously expected, and the validity and reliability are affected by factors like text, starting point, and deletion rate (e.g., Grotjahn et al., 2002). Consequently, C-Tests were developed as a version that takes this criticism into account (e.g., Grotjahn, 1995). Unlike in traditional cloze tests, in C-Tests not entire words are deleted, but parts of words. Through this the test-taker is required to reconstruct the original text by filling in the missing parts of a word (e.g., Grotjahn et al., 2002) which makes the solutions more accessible by having a reduced number of possible acceptable answers and thereby improving scoring efficiency (Grotjahn, 2019). Another difference is that several short texts consisting of approximately 60 to 80 words are used instead of one long text to avoid individual participants gaining an advantage through specialized knowledge in the subject covered in the text (e.g., Grotjahn, 2019). The selected texts should encompass a range of subjects, maintain a neutral stance in terms of content, and be devoid of any specialist terminology (e.g., Grotjahn, 2019).

In classical C-Tests, otherwise referred to as canonical C-Tests, the first sentence is complete in each case and there is also a segment at the end of each text without gaps, which serve as context for finding a solution. Beginning with the second word of the second sentence, the second half of each second word is deleted. The test is therefore based on a mechanical extinguishing principle (e.g., Grotjahn, 2019). An exception is made for proper names and words with a single letter, as these are ignored. If the word to be deleted has an odd number of letters, one letter more is deleted in each case (e.g., Grotjahn, 1995, 2019; Grotjahn et al., 2002). By standardizing the gaps like described, the influence of the starting point and deletion rate on the test's measurement properties is minimized in comparison to cloze tests (Grotjahn, 2019). Through the applied deletion pattern, a high number of test items are constructed in a short text. The typical test comprises four to six texts in total, each containing between 20 and 25 gaps. To facilitate psychometric analysis, it is imperative that all texts contain the same number of gaps (Grotjahn, 2019). The texts are organized from the easiest to the most difficult (e.g., Grotjahn, 2002) and it is recommended that students are allocated between 4 and 6 min to complete each text (e.g., Grotjahn, 2019).

2.2 Measurement and application of C-Tests

C-Tests are regarded as a comprehensive approach for evaluating general language proficiency in first, second, and foreign languages. This general language proficiency is typically conceptualized as a one-dimensional construct, which can be decomposed into microstructural components, and is posited to underpin all language performance, including the classic four language skills (e.g., Grotjahn, 2019). Assuming the assessment of all four language skills lead to discussion in the past as C-Tests do not engage students in oral skills such as listening and speaking (e.g., Alderson, 2002; Shohamy, 1982), which is why some studies argue that these sub-skills should be tested using separate test formats (e.g., Baur et al., 2006). A study by Baghaei and Grotjahn (2014) argues that C-Tests based on spoken discourse may better assess students' listening and speaking skills than C-Tests that rely solely on written discourse texts. Still, numerous studies have demonstrated high correlations between C-Test scores and scores in all four classical language skills. This suggests that C-Tests measure general language proficiency as an integrative construct encompassing all four skills and can therefore be considered a highly reliable, one-dimensional instrument (e.g., Asano, 2014; Eckes and Grotjahn, 2006).

Due to its language evaluating function, C-Tests are primarily employed for the efficient assessment of general language proficiency, for example in the context of placement decisions. They frequently serve as rapid screening tools, preceding more comprehensive test procedures. Additionally, they are utilized for the determination of language proficiency in language acquisition studies and in educational monitoring studies (e.g., Grotjahn, 1995, 2014). In addition, C-Tests are used on meso- and micro-level to make statements about language proficiency by comparing the test score of one class with that of a reference population, through which the tests can take on the status of comparative work (Baur et al., 2006). When re-administering the tests at a later time, it is possible to ascertain both individual progress and the increase in the class average. This, in turn, allows for the derivation of suitable support measures. For further information on the utilization of the tests as support measures, see Baur et al. (2006).

2.3 Evaluation and interpretation of C-Test results

There are many ways of coding and analyzing the test-takers' attempts to reconstruct gap words (Grotjahn, 2019). However, the authors have elected to focus exclusively on the methodology that is pertinent to the remainder of the present paper.

The evaluation process is generally conducted in tabular form. Two distinct values can be ascertained for each individual: first, a correct/incorrect value (referred to as CI-value), which provides information about general language proficiency. A point is awarded if the gap is filled in correctly in terms of semantics, grammar, and orthography. Second, the word-recognition value (referred to as WR-value) which is about receptive language skills and rather focuses on assessing text comprehension. Here, a point is awarded if the word is correctly recognized but an orthographic or grammatical error occurs (Baur and Spettmann, 2010). However, it should be noted that a potential disadvantage of the WR-value is its lack of objectivity, particularly in cases where there is ambiguity. Furthermore, there is a paucity of research on the reliability of the difference scores (e.g., Grotjahn, 2019). Still, at the individual level, the discrepancy between the two scores thus reflects the ratio of receptive and productive language skills (e.g., Baur and Spettmann, 2010; Baur et al., 2013).

2.4 Modifications of C-Test formats

Köberl and Sigott (1994) and Sigott and Köberl (1996) utilized three different modifications of German and English C-Tests in addition to the classical canonical principle and analyzed them with respect to test difficulty. The modifications were that,

(a) two thirds of every second word were deleted,

(b) only the first letter of every second word was shown, and

(c) the first instead of the second half of every second word was deleted.

An experimental design was selected for the study with the deletion patterns being applied to four distinct texts. The findings indicate that in both English and German, best performance is attained with the conventional canonical C-Test, while the least accurate responses are observed when only the initial letter of every second word is provided (b). In German, tests where only a third of every second word is given are solved second best (a), and in English variation (c) was solved second best. One potential explanation for this is that in the English language, deleting the initial half of a word results in a reduced loss of information when compared to German (Sigott and Köberl, 1996). This finding indicates that an increase in redundancy reduction results in C-Tests becoming more challenging, and the implementation of such modifications can be regarded as a potential alternative when the conventional canonical C-Test is deemed to be insufficiently challenging (Sigott and Köberl, 1996).

A study by Lee et al. (2019) confirms that the strategies used to determine which words should be turned into gaps as well as adjustments to the gap size can effectively influence the difficulty of C-Tests. The analysis revealed statistically significant effects on both the participants' error rates and their perceived difficulty. In addition to increasing the difficulty, the procedure aims to measure the degree of knowledge automation and the efficiency of information processing to a greater extent than conventional C-Tests, going beyond the scope of declarative knowledge (Grotjahn, 2010; Wockenfuß, 2009). Ways of modifying the tests to make them easier is done, for example, in the area of people with German as a second language, by Dürrstein (2013). Here, the deletion is confined to the second half of every third word. Nevertheless, this approach has the potential to engender complications, including a diminution in the extent of the psychometrically problematic dependence between the gaps. One method of facilitating the process of testing is to draw a line for each missing letter. However, this can also have a negative effect on validity (Grotjahn, 2019). Another way for modifying tests is to use discourse-specific C-Tests, which are designed to address a particular topic or phenomenon. In this approach, the selection of the gap does not adhere to a mechanical but rather a contextual principle (Baur et al., 2006).

It is important to note that a modified C-Test evaluates different skills compared to a typical canonical C-Test (Grotjahn, 2002). General vocabulary as well as specialized lexical knowledge can be tested, for example, by deleting the first half of every second to third word (e.g., Höttecke et al., 2017), as this approach demands a more advanced level of vocabulary knowledge and the deployment of strategies for constructing meaning (e.g., Baur et al., 2006; Baur and Spettmann, 2010). The target area-orientated selection of those lexemes that are considered to be particularly subject-specific and can further shift the construct from measuring general language proficiency to subject-specific language proficiency (Höttecke et al., 2017). In inflected languages such as German, the deletion of the first half results in an increased focus on macro-structural lexical-semantic skills, thereby reducing the significance of productive knowledge of inflectional morphology during the process of solving (e.g., Beinborn, 2016; Mashkovskaya and Baur, 2016). Mashkovskaya (2013) posits in her study that C-Tests which have gaps at the beginning of words tend to assess receptive reading skills. However, further research is required to ascertain the appropriate circumstances for the utilization of C-Tests with the gap at the beginning of the word (Baur et al., 2006). Daller (1999) moreover found that C-Tests dealing with a newspaper article and C-Tests dealing with academic texts measure two different aspects of language proficiency, referred to as everyday language proficiency and academic language proficiency. This means that an alternative approach to evaluating specialized language proficiency involves the utilization of discourse-specific C-Tests. In this context, it is necessary to assume that the tested individual is familiar with the content of the text; otherwise, technical terms cannot be deduced from the text (Baur et al., 2006).

2.5 Gender differences in solving C-Tests

Various studies have demonstrated significant disparities in test scores between male and female students within educational institutions. Boys have been shown to outperform girls in mathematics and science, while girls have typically achieved higher scores in reading and literacy (e.g., Deutsches PISA-Konsortium, 2000; DESI-Konsortium, 2008; Organisation for Economic Co-operation Development, 2013). Grotjahn et al. (2002) compared C-Test results in French, German, Spanish, English, and Russian. They found that female participants, except for one sample, outperformed their male counterparts. The precise language background of the test-takers remains ambiguous, as the data does not specify whether the tests were taken in the first, second, or foreign language. This finding is corroborated by Mashkovskaya (2013), who surveyed German primary school teacher training students and employed C-Tests in their first language. The study revealed that female participants demonstrated a significantly higher level of performance when compared to males. Höttecke et al. (2017) conducted a study in which they examined the performance of test-takers in both general and discourse-specific C-Tests in the domains of physics and sport. These tests were administered in the test-taker's first language. The findings revealed that female participants demonstrated significantly higher performance in general C-Tests ($d = 0.50$), while their performance in discourse-specific C-Tests did not differ significantly from that of their male counterparts. A study of Iranian intermediate EFL-learners revealed no significant impact of gender on C-Test performance (Tabatabaei and Shakerin, 2013).

2.6 Empirical findings of the relationship between language and mathematical performance

Several empirical studies have demonstrated a correlation between general language and mathematical performance (e.g., Fuchs et al., 2006; Mücke, 2007; Paetsch et al., 2015; Ufer and Bochnik, 2020; Ufer et al., 2013). In the majority of cases, reading skills or an overall score comprising various language skills were included in the analysis of the studies. However,

there are only a few empirical results on which specific general language skills are related to mathematical skills that allow for clear interpretation by including the linguistic sub-competences in the analyses in a differentiated way, simultaneously and under the control of general cognitive abilities. Existing findings that do so are inconclusive. For instance, Beal et al. (2010) demonstrated that reading comprehension emerged as the sole significant predictor of text task solution for second-language learners, while listening comprehension, oral language proficiency, and writing skills did not contribute to this prediction. In contrast, Kempert et al. (2011) found that only oral language proficiency proved to be a significant predictor for solving word problems in multilingual students. For a comprehensive overview of extant empirical studies on different facettes of language proficiency and mathematics, see Paetsch (2016).

The strongest effects were found for the correlation between reading and mathematical skills with a partial correlation of $r = 0.63$ between both domains after statistical control for cognitive abilities (Leutner et al., 2004). Peng et al. (2020) conducted a meta-analysis of the relationship between language and mathematical skills. This analysis, which encompassed 344 studies with 393 independent samples and over 360,000 participants, found a correlation of $r = 0.42$ between the two domains. Particularly the role of text comprehension in solving realistic text tasks has been widely researched and its significant role has been confirmed (e.g., Leiss et al., 2019; Klotz et al., 2025). However, it is not only performance in language demanding text tasks that is linked to language proficiency, but also in tasks that use little language, e.g., arithmetic tasks (e.g., Berg, 2008; Fuchs et al., 2008; Viljaranta et al., 2009). Nevertheless, the impact of language is considerably diminished in tasks that demand minimal language proficiency in comparison to those that are predominantly text-based (e.g., Prediger et al., 2018). A study by Brunner et al. (2022) examined the relationship between the mathematical content areas of arithmetic, word problems, and geometry and general language proficiency measured using typical canonical C-Tests, which covered general topics. The sample consisted of third-graders. The findings of the study demonstrate that there are significant correlations with performance in mathematical tasks. Significant correlations were found between C-Test performance and word problem tasks ($r = 0.500$, $p < 0.001$) but also with relatively poor-in-language geometric tasks ($r = 0.407$, $p < 0.001$). The weakest yet highly significant correlation was found with arithmetic tasks ($r = 0.314$, $p < 0.001$).

Studies found that intercorrelation between language and mathematical skills occur at all ages and levels of education (e.g., Duarte et al., 2011; Greisen et al., 2021; Gürsoy et al., 2013; Kempert et al., 2011; Leiss et al., 2019; Paetsch et al., 2016; Plath and Leiss, 2018; Prediger et al., 2015; Viesel-Nordmeyer et al., 2020), and independent of students' first language (e.g., Ufer et al., 2020).

2.7 The role of subject-specific language skills for mathematical performance

Research has demonstrated that subject-specific language plays a pivotal role in the learning of mathematics. Several studies have been conducted about the relationship between general and mathematical vocabulary in young children and their mathematical performance. These studies include those by Purpura and Reid (2016) and Toll and Van Luit (2014). Utilizing disparate research designs, both studies ascertained that general vocabulary predicted mathematical performance in young children, when mathematical vocabulary was excluded from the models. However, following the incorporation of mathematical vocabulary, the predictive value of general vocabulary was rendered moot. Ufer and Bochnik (2020) found that subject-specific language skills from elementary school students demonstrated a significant relationship to mathematical learning in arithmetic gain again beyond general language skills and other control variables. It shows that having a certain command of a subject-specific school register has an additional impact on students' mathematical learning progress. This corroborates earlier research by Bochnik (2017) who also established that subject-specific language skills function as mediators between general language proficiency and mathematical skills. Here, in particular text-integrative comprehension has been demonstrated to be a more significant predictor of mathematical performance than purely subject-specific vocabulary (Bochnik and Ufer, 2016). Peng and Lin (2019) also identified a correlation between subject-specific language skills and solving text tasks.

3 Materials and methods

3.1 Sample

To answer the research questions, 190 seventh-graders (45.3 % female, 54.7 % male) from 8 classes were surveyed. The participating students were distributed across an integrative comprehensive school (71.1 %) and a lower secondary school track (28.9 %), with an average age of 12.91 years ($SD = 0.61$).

3.2 Study design

The present study constitutes a cross-sectional design, wherein students undergo C-Tests and the mathematical test DEMAT6+ (German Mathematics Test for Sixth Graders, German: Deutscher Mathematiktest für sechste Klassen). In the C-Tests, the characteristics gap placement (beginning vs. end) and text topic (mathematical vs. general) were systematically varied in the sample, and two different evaluation methods (CI vs. WR) were applied.

The test was administered by a trained administrator who had received prior training and had been provided with instructions on how to set the tasks. This ensured that the tests were carried out in an objective manner. For the DEMAT6+, the manual comes with instructions to provide this objectivity (Götz et al., 2013). Following a brief general introduction, the DEMAT6+ and the test booklets were distributed to the participants for the administration of the test, which was conducted in the format of a paper-pencil test. The personal data of the participants, including their age and gender, was collected from the cover page of the test booklet.

A total of 31 min was allotted for the DEMAT6+ examination, with 11 min allocated for the arithmetic part, 5 min for the

geometry part, and 15 min for the word problem part, as outlined in the test manual (Götz et al., 2013). To ensure synchronization of the pages, a stopwatch was utilized to record the elapsed time and to provide a signal to indicate the turning of the pages. Following the completion of the mathematical test, the students proceeded to work on the C-Tests. In this study, the students were allotted a total of 35 min to complete the seven C-Tests, which equates to 5 min per text.

3.3 Description of DEMAT6+

The DEMAT6+ was utilized to assess mathematical performance. The test was developed based on the curricula and educational standards in mathematics and thus considers the guiding principles of "number and operation", "space and form", "size and measurement", "functional context", and "data and chance". The test is to be administered between 6 weeks before the end of the sixth grade and the end of the first semester of the seventh grade, thus rendering is suitable for the selected sample. The test covers three content areas—arithmetic, geometry, and word problems—consisting of 31 tasks that measure mathematical performance in a highly reliable ($\alpha = 0.92$) and valid way (Götz et al., 2013). This is an economical test procedure that is suitable for use in large samples in a research context (Götz et al., 2013). The test was carried out as described in the manual and is fully standardized.

The arithmetic test part (16 items) encompasses questions pertaining to the fundamental concepts of the number system of fractions, conversion between fractions and decimals, and handling units of measurement. Furthermore, the test examines the application of arithmetic rules and the formation and transformation of terms. The geometry test part (4 item) comprises tasks that require the measurement of two-dimensional geometric figures, the understanding of symmetry, and the calculating of the perimeters, areas, and volumes of geometric bodies. The word problem test part (11 items) involves tasks in which information must be extracted from factual contexts and linked to develop a solution. Furthermore, the processing and interpretation of tables and diagrams is examined (Götz et al., 2013).

The manual for the DEMAT6+ contains sample solutions. An answer is classified as either incorrect according to the sample solution (scored with 0) or as one that corresponds to the sample solution (scored with 1).

3.4 Description of C-Test modifications

In total, each student worked on seven texts (word count: $M = 90.4$) which were arranged in ascending order of difficulty. A total of five texts addressed mathematical topics (texts 1–5), while two addressed general topics (texts 6–7). Each text had a total of 25 gaps. Starting with the second sentence, parts of every second word were deleted. At the integrated comprehensive school, always the second half of a word was deleted. At the lower secondary school track, the first half of a word was deleted in three texts instead of the second half to determine differences in difficulty not only between text topic but also between gap placement. It is important to note that the texts used were in German, as the test was administered in German schools. Not all German syntactic constructions can be formed analogously to English. Consequently, a one-to-one translation in terms of gap placement is not feasible.

Each text was evaluated individually. As proposed by Baur and Spettmann (2010), two distinct evaluation methods were conducted (see *Evaluation and interpretation of C-Test results*). For the CI-value, gap fillings are classified as correct and assigned a score of 1 when the gap is filled in a linguistically accurate manner in terms of orthography and grammar. This indicates that the answer is completely correct. For the WR-value, gap fillings are classified as correct and assigned a score of 1 when the meaning of the word is correctly recognized, but a spelling or grammatical error is made. Incorrect or missing answers are assigned a score of 0 in both cases. The present study employs the distinction to ascertain whether the relationship between C-Tests and mathematical performance varies according to the conceptualization of language proficiency as understanding a text and recognizing the word, or in terms of semantic and grammatical correctness. Consequently, conclusions about the relationship between receptive and productive language skills and mathematical performance can be drawn.

3.4.1 Mathematical texts

The construction of mathematical discourse-specific C-Tests necessitated the selection of texts that exhibit semantic and syntactic characteristics typical of the subject under discussion. An exemplar of such a text could be a text from a textbook (Höttecke et al., 2017).

A selection of mathematical topics from various content areas were chosen for the five texts: the subtraction of fractions (text 1), the definition of a quantity of numbers (text 2), the rounding of decimal numbers (text 3), the description of the rule of three (text 4), and a construction instruction (text 5). All texts except text 1 are taken from the schoolbook *MatheNetz 6*. The subtraction is derived from the schoolbook *Faktor 6*. Table 1 illustrates exemplary test items from text 1 with the gap at the end of the word and with the gap at the beginning of the word. A translation is provided.

3.4.2 General texts

The general texts addressed two topics. First, smoking bans on trains and in public offices (text 6). Second, the phenomenon of individuals going to work despite being sick (text 7). The texts are drawn from the DCLL+3, a standardized test for grades 7 and 8, which can be obtained through the *Test Development and Diagnostics Unit of the Hamburg Institute for Educational Monitoring and Quality Assurance*. Table 2 illustrates exemplary test items from text 6 with the gap located at the end of the word and at the beginning of the word. A translation is provided.

3.5 Data analysis

Generalized Linear Mixed Models (GLMMs) were calculated to address the research questions of this study. The analyses were conducted using the *glmer* function of the *lme4* package in *R*

TABLE 1 Exemplary test items from text 1 with gaps at the end and beginning of the word.

Gap placement	Exemplary test items
End	Gleichname Brü___ besitzen d___ gleichen Nen___, ungleichnamige Brüche besi___ zwei untersch___ Nenner.
Beginning	Gleichnamige ___che besitzen ___n gleichen ___ner, ungleichnamige ___tzen zwei ___iedliche Nenner. *Fractions with the same numerator have the same denominator, fractions with different numerators have two different denominators.*

TABLE 2 Exemplary test items from text 6 with gaps at the end and beginning of the word.

Gap placement	Exemplary test items
End	Auch i___ Arbeitsagenturen u___ rund 500 Minis___, Behörden u___ Gerichten gi___ dann Rauch___.
Beginning	Auch ___n Arbeitsagenturen ___d rund 500 ___erien, Behörden ___d Gerichten ___lt dann ___erbot. *Smoking will then also be banned in employment agencies and around 500 ministries, government agencies, and courts.*

version 4.4.2 (Bates et al., 2015). The GLMMs were estimated using a binomial logistic link function.

For RQ1, the models were employed to examine the influence of specific C-Test characteristics on item difficulty. Each gap within the C-Tests was treated as a separate item, and a long-format dataset was constructed in which item-level characteristics were assigned to each individual gap. In this context, the test characteristics (gap placement, text topic, and evaluation method) served as independent variables, while student performance on each item represented the dependent variable. To ascertain the influence of gender on test performance and whether there are differences for C-Tests with certain characteristics, for RQ2 the gender variable was included in one model as an independent variable as well as for calculating interaction effects between gender and test characteristics in another model. In addition to reporting the estimated beta coefficients, we calculated predicted probabilities for key interactions. Reporting these probabilities alongside the beta coefficients presented in the GLMMs allows for a more intuitive interpretation of the magnitude and direction of effects. For RQ3, GLMMs were utilized to investigate how C-Test characteristics relate to mathematical performance across different content areas. Therefore, in addition to the C-Test characteristics (gap placement, text topic, and evaluation method), the mean scores of mathematical performances in the three content areas as well as the overall mean mathematical score were included as independent variables in the analysis to ascertain whether mathematical performance is related to language performance measured through C-Tests. All mathematical variables were grand-mean centered. Furthermore, interaction effects between mathematical performance in different content areas and C-Test characteristics were calculated to analyze whether a certain language skill measured through a modified C-Test is more related to mathematical performance than another. As for RQ2, in addition to reporting the estimated beta coefficients, we calculated predicted probabilities for key interactions.

Prior to all analyses, multicollinearity of the individual independent variables was tested. Given the consistent finding of correlations below 0.8, the presence of suppression effects in the analyses was not anticipated. To account for variation due to individual and item-specific factors, *task_name* (each C-Test gap) and *pid* were included as random effects in all models.

TABLE 3 Influence of different C-Test characteristics on the correct solution frequency of C-Test items.

Predictors	Model 1		
	Estimate	*Std. error*	*p*
(Intercept)	−3.03***	0.25	<0.001
Evaluation method[a]	−0.60***	0.16	<0.001
Gap placement[b]	2.19***	0.19	<0.001
Text topic[c]	1.01***	0.18	<0.001
Random effects			
σ^2	3.29		
τ_{00}	3.03 $_{task\_name}$		
	2.49 $_{pid}$		
N	190 $_{pid}$		
	500 $_{task\_name}$		
Marginal R^2/conditional R^2	0.088/0.659		
Deviance	55,973.604		
AIC	55,985.604		

[a]0 = WR, 1 = CI, [b]0 = beginning, 1 = end, [c]0 = general, 1 = mathematical, ***p < 0.001.

4 Results

Results concerning RQ1: how do different characteristics of the C-Test affect test performance?

GLMMs were used to investigate the impact of different characteristics on the correct solution frequency in C-Tests (see Table 3). In the first model, the impact of the C-Test characteristics evaluation method, gap placement, and text topic on the number of correctly solved items were analyzed.

In Model 1, all included predictors were found to be highly significant. The use of the CI-method was associated with a significantly lower correct solution rate ($\beta = -0.60$, $p < 0.001$), while items with gaps at the end of words lead to a significantly higher correct solution rate ($\beta = 2.19$, $p < 0.001$). Thematic differences have also been demonstrated to be significant. The findings indicated that items covering mathematical topics were

TABLE 4 Influence of different C-Test characteristics and gender on the correct solution frequency of C-Test items.

Predictors	Model 1			Model 2		
	Estimate	Std. error	*p*	Estimate	Std. error	*p*
(Intercept)	−2.50***	0.27	<0.001	−2.74***	0.28	<0.001
Evaluation method[a]	−0.60***	0.16	<0.001	−0.58***	0.16	<0.001
Gap placement[b]	2.20***	0.19	<0.001	2.57***	0.20	<0.001
Text topic[c]	1.01***	0.18	<0.001	0.83***	0.18	<0.001
Gender[d]	−0.97***	0.22	<0.001	−0.50*	0.25	<0.05
Interaction effects						
Gender x evaluation method				−0.04	0.04	0.350
Gender x gap placement				−0.74***	0.12	<0.001
Gender x text topic				0.33***	0.05	<0.001
Random effects						
σ^2	3.29			3.29		
τ_{00}	$3.03_{\text{task\_name}}$			$3.02_{\text{task\_name}}$		
	2.25_{pid}			2.29_{pid}		
N	190_{pid}			190_{pid}		
	$500_{\text{task\_name}}$			$500_{\text{task\_name}}$		
Marginal R^2/conditional R^2	0.115/0.660			0.112/0.660		
Deviance	55,955.203			55,864.612		
AIC	55,969.203			55,884.612		

[a]0 = WR, 1 = CI, [b]0 = beginning, 1 = end, [c]0 = general, 1 = mathematical, [d]0 = female, 1 = male *p <0.05, ***p <0.001.

solved correctly more often than those covering general topics ($\beta = 1.01$, $p < 0.001$).

Results concerning RQ2: are there differences between girls and boys?

In addition to the fixed effects of the aforementioned C-Test characteristics from RQ1, the personal characteristic of gender on the number of correctly solved items was included in the analysis in Model 1. In the second model, interaction effects were considered between gender and the three C-Test characteristics (see Table 4). Moreover, predicted probabilities were calculated (Table 5).

In Model 1, the influence of the student's gender was additionally considered. In this instance, too, all previous effects from RQ1 remained significant. The gender of the student exhibited a substantial negative effect ($\beta = -0.97$, $p < 0.001$), signifying that girls performed significantly better than boys. In Model 2, the interaction effect between gender and evaluation method revealed that girls and boys scored lower when using the CI-method compared to the WR-method. The predicted probabilities indicated that girls achieved 53 % correct answers when evaluating the C-Tests according to the WR-method and 38 % when evaluating the C-Tests according to the CI-method, whereas boys achieved 31 % correct answers when evaluating the C-Tests according to the WR-method and 19 % when evaluating the C-Tests according to the CI-method, suggesting a similar extent of decrease for both genders, consistent with the non-significant interaction ($\beta = -0.04$, $p = 0.350$). Furthermore, girls and boys demonstrated enhanced performance when the gap was located at the end of the word as opposed to the beginning. Predicted probabilities showed that for girls the probability of a correct response increased from 8 % when the gap was at the beginning to 53 % when the gap was at the end, whereas for boys it increased from 6 % (gap at the beginning) to 29 % (gap at the end). This illustrated the significant interaction between gender and gap position, with the effect notably stronger for girls than for boys ($\beta = -0.74$, $p < 0.001$). Finally, the interaction effect between gender and text topic demonstrated that girls and boys performed better on items covering mathematical

TABLE 5 Predicted probabilities of correctly solving C-Test items with different characteristics by student gender.

C-Test characteristic	Girls	Boys
Evaluation method		
WR-method	53 %	31 %
CI-method	38 %	19 %
Gap placement		
Beginning	8 %	6 %
End	53 %	29 %
Text topic		
General	31 %	12 %
Mathematical	51 %	31 %

Percentages indicate predicted probabilities of correctly solving C-Test items.

topics than on items covering general topics, yet boys benefited significantly more ($\beta = 0.33$, $p < 0.001$). Predicted probabilities indicated that girls achieved 31 % correct answers on items covering general topics and 51 % on items covering mathematical topics, whereas boys achieved 12 % correct answers on items covering general topics and 31 % on items covering mathematical topics.

Results concerning RQ3: how is students' mathematical performance in different content areas related to their C-Test performance, and how is this relationship influenced by the C-Test characteristics gap placement, text topic, and evaluation method?

GLMMs were utilized to examine the impact of different C-Test characteristics as well as mathematical performance in the content areas arithmetic, geometry, and word problems on the correct solution frequency in C-Tests (see Table 6). In Model 1, the effects of the C-Test characteristics evaluation method, gap placement, and text topic as well as overall mathematical performance and the interaction effects between mathematical performance and C-Test characteristics on the number of correctly solved C-Test items were analyzed. In Model 2, the same approach was adopted for the analysis of performance in arithmetic tasks, with the objective of differentiating more precisely between the content areas. In Model 3, analog analyses were conducted for geometry and in Model 4 for word problems. Moreover, predicted probabilities were calculated (Table 7).

In all four models, as was the case in the analysis for RQ1 and RQ2, all C-Test characteristics were found to be highly significant. The employment of the CI-method was associated with a significantly lower correct solution rate, while gaps at the end of words lead to a significantly higher correct solution rate and items covering mathematical topics were solved correctly more often than those covering general topics.

Overall mathematical and C-Test performance indicated a positive relationship, suggesting that higher mathematical performance was associated with higher C-Test performance ($\beta = 4.90$, $p < 0.001$; Model 1) and therefore, consequently, higher language proficiency. Regarding the different content areas, a highly significant relationship has been demonstrated between arithmetic and C-Test performance ($\beta = 4.93$, $p < 0.001$; Model 2), and between performance in word problems and C-Tests ($\beta = 2.80$, $p < 0.001$; Model 4). It was found that a higher level of geometric performance was significantly associated with higher C-Test performance ($\beta = 1.88$, $p = 0.002$; Model 3), albeit to a lesser extent than in the other content areas.

Except for geometric performance ($\beta = 0.23$, $p = 0.038$; Model 3), no significant interaction effect between mathematical performance and the evaluation method was identified. Predicted probabilities indicated that students with low geometric skills (-1 *SD*) had a higher probability of correctly filling in the gaps according to the WR-method (34 %) than the CI-method (21 %), suggesting that they benefited more from the WR-method. In contrast, students with high geometric skills ($+1$ *SD*) also benefited more from the WR-method (41 %) than the CI-method (28 %), but the difference was proportionally smaller, indicating that higher-performing students in geometry were less affected by the choice of evaluation method. This indicated that the influence of students' mathematical skills in geometry on C-Test performance was slightly stronger when evaluating the C-Tests according to the CI-method than the WR-method. For arithmetic and word problems as well as for the overall mathematical performance, the negative effect of the CI-method was therefore independent of students' mathematical performance.

There was a highly significant negative interaction effect between gap placement and overall mathematical performance ($\beta = -2.19$, $p < 0.001$; Model 1). Predicted probabilities illustrated that students with low overall mathematical skills (-1 *SD*) correctly solved 2 % of items with the gap at the beginning of the word and 22 % of items with the gap at the end, whereas students with high overall mathematical skills ($+1$ *SD*) solved 9 % of items with the gap at the beginning of the word and 42 % of items with the gap at the end of the word correctly. This indicated that although gaps at the end generally facilitated item completion, students with lower mathematical skills benefited disproportionately from this C-Test characteristic, while higher-performing students in mathematics gained even more overall but showed a smaller relative difference between gap positions. The highly significant negative interaction effect between gap placement and arithmetic performance ($\beta = -2.63$, $p < 0.001$; Model 2) again indicated that the advantage of C-Tests with gaps placed at the end of the word dimished as mathematical skills in this content area increased. Predicted probabilities illustrated this pattern: when the gap was at the beginning of the word, the probability of a correct response rose only slightly with increasing arithmetic performance—from 4 % for low-performing students (-1 *SD*) to 13 % for high-performing students ($+1$ *SD*). In contrast, when the gap was at the end of the word, the corresponding probabilities increased from 31 % to 50 %. This pattern was most pronounced in the arithmetic subdomain when considering the content areas separately, followed by geometry ($\beta = -1.24$, $p < 0.001$; Model 3). Predicted probabilities illustrated this effect for students with low geometric skills (-1 *SD*) and high geometric skills ($+1$ *SD*): when the gap was at the beginning of the word, the probability of a correct response was 4 % for low-performing and 7 % for high-performing students; when the gap was at the end of the word, the corresponding probabilities were 40 % (gap at the beginning) and 47 % (gap at the end). This finding was least pronounced for word problems ($\beta = -1.00$, $p < 0.001$; Model 4). Predicted probabilities indicated that students with low word problem skills (-1 *SD*) correctly solved 2 % of items with the gap at the beginning of the word and 22 % of items with the gap at the end, whereas students with high word problem skills ($+1$ *SD*) correctly solved 6 % (gap at the beginning) and 34 % (gap at the end) of items.

Finally, the analysis demonstrated that students performed better in the C-Tests covering mathematical text topics when they possessed higher overall mathematical skills ($\beta = 1.92$, $p < 0.001$; Model 1). Predicted probabilities indicated that for students with low overall mathematical skills (-1 *SD*), the probability of correctly solving C-Test items was 12 % for items covering general topics and 20 % for items covering mathematical topics, whereas for high-performing students ($+1$ *SD*), the corresponding probabilities were 21 % (general text topics) and 43 % (mathematical text topics), respectively, highlighting that mathematical C-Test items particularly favored students with higher overall mathematical skills. In this case as well, the effect was most pronounced for arithmetic performance ($\beta = 1.64$, $p < 0.001$; Model 2), where predicted probabilities indicated that students with low arithmetic skills (-1 *SD*) solved 15 % of the general C-Test items and 29 %

TABLE 6 Influence of different C-Test characteristics and mathematical performance on the correct solution frequency of C-Test items.

Predictors	Model 1			Model 2			Model 3			Model 4		
	Estimate	Std. Error	*p*	Estimate	Std. Error	*p*	Estimate	Std. Error	*p*	Estimate	Std. Error	*p*
(Intercept)	−2.92***	0.25	<0.001	−2.98***	0.25	<0.001	−3.01***	0.25	<0.001	−2.93***	0.25	<0.001
Evaluation method[a]	−0.60***	0.16	<0.001	−0.60***	0.16	<0.001	−0.60***	0.16	<0.001	−0.60***	0.16	<0.001
Gap placement[b]	2.08***	0.19	<0.001	2.09***	0.19	<0.001	2.17***	0.19	<0.001	2.09***	0.19	<0.001
Text topic[c]	1.01***	0.18	<0.001	1.02***	0.18	<0.001	1.00***	0.18	<0.001	1.02***	0.18	<0.001
Overall mathematical performance (full-scale)[d]	4.90***	0.96	<0.001									
Mathematical performance (full-scale) x evaluation method	0.04	0.18	0.802									
Mathematical performance (full-scale) x gap placement	−2.19***	0.45	<0.001									
Mathematical performance (full-scale) x text topic	1.92***	0.19	<0.001									
Arithmetic performance (sub-scale)[d]				4.93***	1.00	<0.001						
Arithmetic performance (sub-scale) x evaluation method				−0.05	0.19	0.775						
Arithmetic performance (sub-scale) x gap placement				−2.63***	0.52	<0.001						
Arithmetic performance (sub-scale) x text topic				1.64***	0.20	<0.001						
Geometric performance (sub-scale)[d]							1.88**	0.60	0.002			
Geometric performance (sub-scale) x evaluation method							0.23*	0.11	0.038			
Geometric performance (sub-scale) x gap placement							−1.24***	0.29	<0.001			
Geometric performance (sub-scale) x text topic							0.59***	0.12	<0.001			
Word problem performance (sub-scale)[d]										2.80***	0.65	<0.001
Word problem performance (sub-scale) x evaluation method										−0.01	0.12	0.928
Word problem performance (sub-scale) x gap placement										−1.00***	0.30	0.001
Word problem performance (sub-scale) x text topic										1.11***	0.13	<0.001

(Continued)

TABLE 6 (Continued)

Predictors	Model 1			Model 2			Model 3			Model 4		
	Estimate	Std. Error	*p*	Estimate	Std. Error	*p*	Estimate	Std. Error	*p*	Estimate	Std. Error	*p*
Random effects												
σ^2	3.29			3.29			3.29			3.29		
τ_{00}	3.04 $_{task\_name}$			3.05 $_{task\_name}$			3.03 $_{task\_name}$			3.04 $_{task\_name}$		
	2.24 $_{pid}$			2.34 $_{pid}$			2.43 $_{pid}$			2.26 $_{pid}$		
N	190 $_{pid}$			190 $_{pid}$			190 $_{pid}$			190 $_{pid}$		
	500 $_{task\_name}$			500 $_{task\_name}$			500 $_{task\_name}$			500 $_{task\_name}$		
Marginal R^2/Conditional R^2	0.130/0.666			0.113/0.663			0.098/0.661			0.127/0.666		
Deviance	55,829.285			55,871.112			55,920.201			55,866.130		
AIC	55,849.285			55,891.112			55,940.201			55,886.130		

[a]0 = WR, 1 = CI, [b]0 = beginning, 1 = end, [c]0 = general, 1 = mathematical, [d]mean score of test tasks divided by number of test tasks. *p <0.05, **p <0.01, ***p <0.001.

TABLE 7 Predicted probabilities of correctly solving C-Test items for low- and high-achieving students in mathematics across content areas for significant interaction effects.

Content area	C-Test characteristic	Low (1 *SD* below the mean)	High (1 *SD* above the mean)
Overall performance	Text topic (geneal/mathematical)	12 %/20 %	21 %/43 %
	Gap placement (beginning/end)	2 %/22 %	9 %/42 %
Arithmetic	Text topic (general/mathematical)	15 %/29 %	25 %/52 %
	Gap placement (beginning/end)	4 %/31 %	13 %/50 %
Geometry	Text topic (general/mathematical)	17 %/33 %	20 %/41 %
	Evaluation method (WR/CI)	34 %/21 %	41 %/28 %
	Gap placement (beginning/end)	4 %/40 %	7 %/47 %
Word problems	Text topic (general/mathematical)	12 %/20 %	17 %/34 %
	Gap placement (beginning/end)	2 %/22 %	6 %/34 %

Percentages indicate predicted probabilities of correctly solving C-Test items.

of the mathematical C-Test items correctly, whereas students with high mathematical skills (+1 *SD*) solved 25 % (general text topics) and 52 % (mathematical text topics) of these items correctly. This finding was second-most pronounced for word problems ($\beta = 1.11$, $p < 0.001$; Model 4). Predicted probabilities showed that students with low word problem skills (-1 *SD*) correctly solved 12 % of items with general topics and 20 % of items with mathematical topics, whereas students with high word problem skills (+1 *SD*) correctly solved 17 % (general text topics) and 34 % (mathematical text topics) of these items. Finally, this finding was least for geometric performance ($\beta = 0.59$, $p < 0.00$; Model 3). Here, predicted probabilities indicated that students with low geometric skills (−1 *SD*) correctly solved 17 % of items with general topics and 33 % of items with mathematical topics, whereas students with high geometric skills (+1 *SD*) correctly solved 20 % (general text topics) and 41 % (mathematical text topics) of items, still showing that mathematical items favored higher-performing students, but the effect was smaller than for arithmetic and word problems.

5 Discussion

In our study, first the effects of the C-Test characteristics gap placement, text topic, and evaluation method as well as the sociodemographic variable gender on test performance were investigated. Furthermore, interaction effects between gender and C-Test characteristics were analyzed. Second, the study examined the relationship between language skills measured through C-Tests and overall mathematical performance as well as mathematical performance in different content areas. Moreover, interaction effects between the different C-Test characteristics and mathematical performance were analyzed to further exploit the linkage between language and mathematical skills and at the same time drive new insights into the diagnostic and educational potential of C-Tests.

The analysis for RQ1 (*How do different characteristics of the C-Test affect test performance?)* shows that formal characteristics of C-Tests have a highly significant influence on the correct solution rate. Particularly noteworthy is the strong effect of the gap placement position, which indicates that C-Tests with the gap at the end of the word are solved significantly better than C-Tests with the gap at the beginning of the word. This confirms previous findings (e.g., Baur and Spettmann, 2010), who further concluded that besides necessary higher vocabulary knowledge, a reason for the increased difficulty could be that C-Tests with the gap at the beginning of the word are more likely to challenge test-takers to activate specific cognitive strategies to construct meaning when filling in the gaps, which could not be present in the students of this study. Moreover,

the analysis confirms that C-Tests covering mathematical text topics are solved significantly better than C-Tests covering general text topics. This shows that basic mathematical knowledge is present in the students and that subject-specific language elements can be used effectively. These results indicate that the participants' subject-specific language proficiency is better that their general language proficiency. Assumptions can be made as to why this is the case. There may be too few general texts, which can potentially create a bias, as the findings only relate to two specific text topics and do not allow a broader statement. Studies showed that the average score on any C-Test is known to differ due to topic familiarity (Kamimoto, 1993) and textual types (Mochizuki, 1994), indicating that students could not be familiar enough with the general topics and text type covered in the chosen general C-Tests and are more familiar with the mathematical text type and topic, for instance by working with mathematical text tasks often. The mathematical part consisted of five texts that covered a wide variety of topics and therefore addressed different aspects of mathematical knowledge, which may allow more points to be scored across the texts on average. Nevertheless, it is still possible that students have a more developed mathematical vocabulary than general vocabulary. To make a clear statement about this, further analysis regarding the texts (e.g., knowledge about the topic, motivational factors) are recommended. Lastly, using the WR-method when evaluating C-Tests leads to significantly better test results. This finding is not surprising because here answers are more often counted as correct than using the CI-method. Still, when choosing only one evaluation method, the impact it can make on the determination of the test-taker's language proficiency must be kept in mind. H1 can therefore be confirmed, with the exception that C-Tests covering mathematical text topics were solved more successfully than C-Tests covering general text topics.

For RQ2 (*Are there differences between girls and boys?)* the analysis reveals that, on average, girls perform significantly better than boys in C-Tests, which supports findings from Grotjahn et al. (2002) and Mashkovskaya (2013) and generally confirms previous realizations that girls score higher in literacy (e.g., Organisation for Economic Co-operation Development, 2013). Yet, five out of seven texts were related to mathematical topics, suggesting that the linguistic demands of the test might have a stronger influence on performance than the mathematical content. This could indicate that C-Tests primarily assess linguistic rather than thematic—here mathematical—knowledge. Still, the interaction analysis reveals that boys benefit more from C-Tests with mathematical than general topics in comparison to girls, which is again showing the strong performance of boys in mathematics, even when it is not about solving mathematical tasks but texts dealing with mathematical topics. The interaction analysis moreover reveals that girls benefit more from C-Tests where the gap is at the end of the word than boys. One potential explanation for this phenomenon is that girls tend to exhibit a stronger metalinguistic awareness in general (Al-Ahdal and Almarshedi, 2021), which seems to be particularly evident in domains necessitated by C-Tests involving the identification of gaps at the end of words. It is evident that the initial gap is addressed less by lexical semantic skills (e.g., Beinborn, 2016) and more by morphological and syntactic skills. This is because the missing word frequently contains the word stem or the complete lexical core. These skills are shown to be particularly present in girls. Lastly, no significant difference between girls and boys can be found according to the chosen evaluation method. The assumptions formulated in H2—that girls generally outperform boys, while boys benefit more from mathematical texts—can be confirmed. Moreover, an additional finding emerged, indicating that girls benefit more strongly from gaps at the end of words than boys.

This first part of the study provides evidence that the characteristics of C-Tests influence test performance and that differences between genders are real. This can therefore influence the language proficiency that is actually measured and attributed to the test-takers, which is why the choice of the gap placement, text topic, and evaluation method should not be arbitrary.

Regarding RQ3 (*How is students' mathematical performance in different content areas related to their C-Test performance, and how is this relationship influenced by the C-Test characteristics gap placement, text topic, and evaluation method?*), it can generally be said that mathematical skills are highly significantly related to C-Test performance and thus with the students' language proficiency. However, there are differences in mathematical content areas. This relationship is strongest for arithmetic, followed by word problems, and weakest for geometry. The strong relationship with word problems also confirms findings from the study from Brunner et al. (2022) and can potentially be explained due to the text-heaviness of the tasks. Yet, it is remarkable that in our study performance in arithmetics is stronger related to C-Test performance that text-heavy word problems, underscoring the importance of language skills for pure arithmetic as also shown in previous studies (e.g., Fuchs et al., 2008). Berg (2008) concluded that despite drawing on different types of knowledge—such as phonological decoding in reading and numerical understanding in arithmetic—both domains still engage comparable cognitive procedures. However, this finding of ours provides new approaches for when problems arise in calculating arithmetic tasks, namely by focusing on language components of the task and the language skills of the test-taker. The fact that geometric tasks are least related to C-Test performance is potentially indicating the need for other skills (e.g., spatial orientation) (Jablonski and Ludwig, 2023) when solving them, which are not detected by any of the C-Test variations. This difference from findings by Brunner et al. (2022), in which the correlation with geometry was second highest, could be explained by different grades.

Furthermore, the analysis shows that students who have higher mathematical skills benefit more from C-Tests with mathematical text topics, indicating that these C-Tests are more closely related to mathematical performance than C-Tests with general text topics, which reinforces the importance of subject-specific language skills. This finding applies to all content areas but is again most pronounced in arithmetic. This supports the findings from previous research, showing that mathematical vocabulary predicts mathematic skills better than general vocabulary, as shown in studies by Purpura and Reid (2016) and Toll and Van Luit (2014). Bochnik (2017) even confirmed that subject related language

acts as a mediator for general language proficiency. This means that understanding specialized mathematical language can have a positive influence on mathematical performance and shows the importance of specialized language skills besides general language skills for mathematical success. Moreover, the analysis reveals that low performing students in mathematics significantly benefit from C-Tests where the gap is at the end of the word, whereas this effect diminishes as mathematical skills increase. Once again, the effect is strongest for arithmetic, followed by geometry, and then word problems. Previous research has shown that deleting the first half of words tends to increase test difficulty by requiring a higher degree of activation of strategies for constructing meaning (e.g., Baur and Spettmann, 2010; Baur et al., 2006). Thus, this increased complexity appears to enhance the relation to mathematical performance as lower achieving students in mathematics significantly benefit from the "easier" C-Tests. This may be explained by the cognitive demands associated with reconstructing word beginnings, which require strategic use of context, inference, and advanced lexical-semantic processing-skills that are also central to solving mathematical tasks. Lastly, the finding that only higher-performing students in geometry are less affected by the choice of evaluation method is remarkable as the influence of students' mathematical skills in geometry on C-Test performance is slightly stronger when evaluating them according to the CI-method than the WR-method. Achieving higher scores when evaluating C-Tests according to the CI-method is associated with higher productive language skills (Baur and Spettmann, 2010; Baur et al., 2013). A possible explanation could be that students with strong productive language skills possibly express their thought processes more clearly, which is important when solving geometric tasks. Yet, more research to clarify this is necessary. H3 can therefore not be confirmed, with the exception that mathematical texts show a stronger relationship with mathematical performance than general texts.

The second part of this study carries significant implications for educational research and practice, particularly with respect to the interplay of language and mathematics education. The results highlight the importance of acknowledging and addressing the language demands inherent in mathematical learning. From a research and education policy perspective, the study reinforces the need for interdisciplinary approaches that connect language and mathematics education. For classroom practice, the study advocates for a more language-sensitive approach to mathematical instructions, which is not only important for text-heavy tasks. Teachers should therefore be supported in fostering students' ability to understand mathematical tasks by integrating language in the mathematics classroom, which could be achieved through interdisciplinary teaching or the use of adequate methods which future research should focus on. Focusing on the potential applications of C-Tests, the findings of this study suggest that to support the development of mathematical skills via C-Tests in the classroom, it is particularly effective to use C-Tests with mathematical text topics. For students with lower mathematical skills, placing the gap at the end of the word enhances performance and provides accessible learning opportunities, whereas higher-achieving students benefit from mathematical C-Tests regardless of gap placement, though the relative advantage of end-gaps is smaller. Here, the gap at the beginning could further increase the challenge. Given their strong association with mathematical performance—particularly in arithmetic—implementing C-Tests on specific, generally less language-intensive topics could enhance students' understanding of the subject matter and improve their mathematical achievement. This then opens promising avenues for identifying learning difficulties and for designing targeted interventions in mathematics lessons based on students' language proficiency. Further research is required to explore this potential in greater depth. Regarding performance differences in language and mathematics subjects, this could moreover be a step toward closing the gender gap in academic achievement between both domains.

6 Limitations and recommendations for further research

The present study is subject to certain limitations. First, the sample size is small and therefore generalizability may not be guaranteed. Moreover, the number of test items between the mathematical and the general C-Tests is unequal. This imbalance has the potential to compromise the comparability of the results between the two variations. In the design of future studies, it is recommended that the number of texts with different topics be maintained equal, with the aim of preventing test bias. A more balanced study design with respect to the distribution of the C-Test items in terms of gap placement would be desirable in future research to determine whether this might reveal additional differences. Furthermore, a more in-depth examination of the specific content of a text and its relationship to the mathematical tasks from different content areas in a future study could potentially yield more precise conclusions. Moreover, it must be considered that German texts were used in this study, and it is unclear how transferable the results are to C-Tests in other languages. C-Tests nowadays are available in more than 30 languages (Grotjahn, 2014). However, originally, they were developed for German and English, which makes a change of the typical deletion pattern necessary for languages that differ greatly from these languages, e.g., Turkish, Japanese, or Chinese (e.g., Grotjahn, 1995, 2002, 2019).

Some studies have also claimed that cognitive processes play a prior role when solving C-Tests. As the present study addressed the language demands of C-Tests, they were not given consideration. However, further studies could ascertain whether modified C-Tests stimulate different cognitive processes that are pivotal to mathematical performance. For instance, Wockenfuß and Raatz (2014) concluded that the performance of native speakers in C-Test depends on processing speed and verbal intelligence and Baghaei and Tabatabaee (2015) argue that the underlying construct of C-Test corresponds with the abilities underlying the language component of crystallized intelligence and through careful text selection also the factual knowledge component of this construct. Lastly, future research could further investigate the influence of students' language-related background characteristics—such as first language or migration background—to enable more differentiated findings.

Data availability statement

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

Ethics statement

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

Author contributions

EK: Visualization, Resources, Project administration, Validation, Formal analysis, Data curation, Investigation, Methodology, Writing – review & editing, Writing – original draft, Conceptualization. TE: Resources, Software, Investigation, Data curation, Conceptualization, Formal analysis, Methodology, Validation, Project administration, Supervision, Writing – review & editing. DL: Supervision, Writing – review & editing, Conceptualization, Investigation, Validation, Resources, Funding acquisition, Project administration.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This publication was funded by the Open Access Publication Fund of Leuphana University Lüneburg.

Acknowledgments

Eileen Klotz would like to express her sincere gratitude to her colleague Pia Carlotta Kohlstedt, who supported her by introducing her to the statistics tool *R* and GLMMs.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

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

Publisher's note

References

Al-Ahdal, A. A. M. H., and Almarshedi, R. M. (2021). Metalinguistic awareness and academic achievement: finding correlations among high-achieving EFL learners. *J. Lang. Linguistic Stud.* 17, 2274–2285.

Alderson, C. J. (2002). "Testing proficiency and achievement: principles and practice," in *University Language Testing and the C-Test*, eds. J. A. Coleman, R. Grotjahn, and U. Raatz (Bochum: AKS-Verlag), 15–30.

Asano, Y. (2014)."C-Tests und, allgemeine Sprachkompetenz': Theoretische Überlegungen und empirische Analysen," in *Der C-Test: Aktuelle Tendenzen*, ed. R. Grotjahn (Frankfurt am Main: Peter Lang), 39–52.

Baghaei, P., and Grotjahn, R. (2014). Establishing the construct validity of conversational C-tests using a multidimensional Rasch model. *Psychol. Test Assess. Model.* 56, 60–82.

Baghaei, P., and Tabatabaee, M. (2015). The C-test: an integrative measure of crystallized intelligence. *J. Intellig.* 3, 46–58. doi: 10.3390/jintelligence3020046

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Baur, R. S., Goggin, M., and Wrede-Jackes, J. (2013). *Der C-Test: Einsatzmöglichkeiten im Bereich DaZ. Universität Duisburg-Essen: proDaZ.* Available online at: http://www.uni-due.de/imperia/md/content/prodaz/c_test_einsatzmoeglichkeiten_daz.pdf (Accessed August 13, 2025).

Baur, R. S., Grotjahn, R., and Spettmann, M. (2006). "Der C-Test als Instrument der Sprachstandserhebung und Sprachförderung," in *Fremdsprachenlernen und Fremdsprachenforschung. Kompetenzen, Standards, Lernformen, Evaluation. Festschrift für Helmut Johannes Vollmer*, ed. J.-P. Timm (Tübingen: Narr), 389–406.

Baur, R. S., and Spettmann, M. (2010). "Lesefertigkeiten testen und fördern," in *Fachliche und sprachliche Förderung von Schülern mit Migrationsgeschichte. Beiträge des Mercator-Symposions im Rahmen des 15. AILA-Weltkongresses "Mehrsprachigkeit: Herausforderungen und Chancen,"* eds. C. Benholz, G. Kniffka, and E. Winters-Ohle (Münster: Waxmann), 97–118.

Beal, C. R., Adams, N. M., and Cohen, P. R. (2010). Reading proficiency and mathematics problem solving by high school English language learners. *Urban Educ.* 45, 58–74. doi: 10.1177/0042085909352143

Beinborn, L. (2016). *Predicting and manipulating the difficulty of text-completion exercises for language learning* (Doctoral dissertation). Technische Universität Darmstadt, Fachbereich Informatik. Available online at: http://tuprints.ulb.tu-darmstadt.de/5647/ (Accessed July 16, 2025).

Berg, D. H. (2008). Working memory and arithmetic calculation in children: the contributory roles of processing speed, short-term memory, and reading. *J. Exp. Child Psychol.* 99, 288–308. doi: 10.1016/j.jecp.2007.12.002

Bochnik, K. (2017). *Sprachbezogene Merkmale als Erklärung für Disparitäten mathematischer Leistung: differenzierte Analysen im Rahmen einer Längsschnittstudie in der dritten Jahrgangsstufe. Empirische Studien zur Didaktik der Mathematik, 30.* Münster; New York: Waxmann.

Bochnik, K., and Ufer, S. (2016). Die Rolle (fach-)sprachlicher Kompetenzen zur Erklärung mathematischer Kompetenzunterschiede zwischen Kindern mit deutscher und nicht-deutscher Familiensprache. *Zeitschrift für Grundschulforschung* 9, 135–147. doi: 10.17877/DE290R-17326

Brunner, E., Bernet, F., and Nänny, S. (2022). Zum Zusammenhang zwischen sprachlichen und mathematischen Leistungen in unterschiedlichen

Inhaltsbereichen. *Swiss J. Educ. Res.* 44, 167–179. doi: 10.24452/sjer.44.2.1

Daller, H. (1999). The language proficiency of Turkish returnees from Germany: an empirical investigation of academic and everyday language proficiency. *Lang. Cult. Curriculum* 12, 156–172. doi: 10.1080/07908319908666575

DESI-Konsortium (2008). "Unterricht und Kompetenzerwerb," in *Zentrale Ergebnisse der Studie Deutsch-Englisch-Schülerleistungen International*. Frankfurt am Main: DIPF.

Deutsches PISA-Konsortium (2000). "PISA 2000," in *Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.

Duarte, J., Gogolin, I., and Kaiser, G. (2011). "Sprachlich bedingte Schwierigkeiten von mehrsprachigen Schüle-rinnen und Schülern bei Textaufgaben," in *Mathematik unter Bedingungen der Mehrsprachigkeit*, eds. S. Prediger, and E. Özdil (Münster: Waxmann), 35–53.

Dürrstein, B. (2013). "Konzeption eines C-Tests als Screening," in *Sprachförderung und Förderdiagnostik in der Sekundarstufe I*, ed. S. Jeuk (Stuttgart: Klett), 53–71.

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

Fuchs, L. S., Compton, D. L., Fuchs, D., Hollenbeck, K. N., Carddock, C. F., and Hamlett, C. L. (2008). Dynamic assessment of algebraic learning in predicting third graders' development of mathematical problem solving. *J. Educ. Psychol.* 100, 829–850. doi: 10.1037/a0012657

Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M., et al. (2006). The cognitive correlates of third-grade skill in arithmetic, algorithmic computation, and arithmetic word problems. *J. Educ. Psychol.* 98, 29–43. doi: 10.1037/0022-0663.98.1.29

Götz, L., Lingel, K., and Schneider, W. (2013). *DEMAT 6+. Deutscher Mathematiktest für sechste Klassen. 1. Auflage*. Göttingen: Hogrefe.

Greisen, M., Georges, C., Hornung, C., Sonnleitner, P., and Schiltz, C. (2021). Learning mathematics with shackles: how lower reading comprehension in the language of mathematics instruction accounts for lower mathematics achievement in speakers of different home languages. *Acta Psychol.* 221, 1–11. doi: 10.1016/j.actpsy.2021.103456

Grotjahn, R. (1995). Der C-test: state of the art. *Zeitschrift für Fremdsprachenforschung* 6, 37–60.

Grotjahn, R. (2002). "Konstruktion und Einsatz von C-Tests: Ein Leitfaden für die Praxis," in *Der C-Test: Theoretische Grundlagen und praktische Anwendungen, Bd. 4*, ed. R. Grotjahn (Bochum: AKS-Verlag), 211–225.

Grotjahn, R. (2010). "Gesamtdarbietung, Einzeltextdarbietung, Zeitbegrenzung und Zeitdruck: Auswirkungen auf Item- und Testkennwerte und C-Test-Konstrukt," in *Der C-Test: Beiträge aus der aktuellen Forschung*, ed. R. Grotjahn (Frankfurt am Main: Peter Lang), 265–296.

Grotjahn, R. (2014). "Der C-Test: Aktuelle Tendenzen. Einleitung und Übersicht über den Band," in *Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends*, ed. R. Grotjahn (Frankfurt am Main: Peter Lang), 7–37.

Grotjahn, R. (2019). "C-Tests," in *Sprachdiagnostik Deutsch als Zweitsprache: ein Handbuch*, eds. S. Jeuk, and J. Settinieri (Berlin: De Gruyter) 585–609.

Grotjahn, R., Klein-Braley, C., and Raatz, U. (2002). "C-Tests: an overview," in *University Language Testing and the C-Test*, eds. J. Coleman, R. Grotjahn, and U. Raatz (Bochum: AKS-Verlag), 43–114.

Gürsoy, E., Benholz, C., Renk, N., Prediger, S., and Büchter, A. (2013). Erlös=Erlösung? Sprachliche und konzeptuelle Hürden in Prüfungsaufgaben zur Mathematik. *Deutsch als Zweitsprache* 1, 14–24.

Höttecke, D., Ehmke, T., Krieger, C., and Kulik, M. A. (2017). Vergleichende Messung fachsprachlicher Fähigkeiten in den Domänen Physik und Sport. *Zeitschrift für Didaktik der Naturwissenschaften (ZfDN)* 23, 53–69. doi: 10.1007/s40573-017-0055-6

Jablonski, S., and Ludwig, M. (2023). Teaching and learning of geometry—a literature review on current developments in theory and practice. *Educ. Sci.* 13:682. doi: 10.3390/educsci13070682

Kamimoto, T. (1993). Tailoring the test to fit the students: Improvement of the C-Test through classical item analysis. *Lang Laborat.* 30, 47–61.

Kempert, S., Saalbach, H., and Hardy, I. (2011). Cognitive benefits and costs of bilingualism in elementary school students: the case of mathematical word problems. *J. Educ. Psychol.* 103, 547–561. doi: 10.1037/a0023619

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: an appraisal. *Lang. Test.* 14, 47–84. doi: 10.1177/026553229701400104

Klotz, E., Ehmke, T., and Leiss, D. (2025). Text comprehension as a mediator in solving mathematical reality-based tasks: the impact of linguistic complexity, cognitive factors, and social background. *Eur. J. Educ. Res.* 14, 23–39. doi: 10.12973/eu-jer.14.1.23

Köberl, J., and Sigott, G. (1994). "Adjusting C-Test Difficulty in German," in *Der C-Test. Theoretische Grundlagen und praktische Anwendungen, Bd. 2*, ed. G. Rüdiger (Bochum: Brockmeyer), 179–192.

Lee, J.-U., Schwan, E., and Meyer, C. M. (2019). "Manipulating the difficulty of C-tests," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Italy: Association for Computational Linguistics), 360–370.

Leiss, D., Plath, J., and Schwippert, K. (2019). Language and mathematics – key factors influencing the comprehension process in reality-based tasks. *Mathemat. Think. Learn.* 21, 201–213. doi: 10.1080/10986065.2019.1570835

Leutner, D., Klieme, E., Meyer, K., and Wirth, J. (2004). "Problemlösen," in *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland? Ergebnisse des zweiten internationalen Vergleichs*, eds. PISA Konsortium Deutschland (Münster: Waxmann), 147–175.

Mashkovskaya, A. (2013). *Der C-Test als Lesetest bei Muttersprachlern* (Diss. phil.). Universität Duisburg-Essen, Fachbereich Geisteswissenschaften. Available online at: http://duepublico.uni-duisburg-essen.de/servlets/DocumentServlet?id=32859 (Accessed September 2, 2025).

Mashkovskaya, A., and Baur, R. S. (2016). "C-Tests für erwachsene Erst- und ZweitsprachsprecherInnen. Screening der (schrift-)sprachlichen Kompetenzen von Lehramtsstudierenden," in *Schriftsprachliche Kompetenzen von Lehramtsstudierenden in der Studieneingangsphase: Eine empirische Untersuchung*, eds. A. Bremerich-Vos, and D. Scholten-Akoun (Baltmannsweiler: Schneider Verlag Hohengehren), 191–211.

Mochizuki, A. (1994). C-Tests: Four kinds of texts, their reliability and validity. *JALT J.* 16, 41–54.

Mücke, S. (2007). "Einfluss personeller Eingangsvoraussetzungen auf Schülerleistungen im Verlauf der Grundschulzeit," in *Qualität von Grundschulunterricht: entwickeln, erfassen und bewerten*, eds. K. Möller, P. Hanke, C. Beinbrech, A. K. Hein, T. Kleickmann, and R. Schages (Wiesbaden: Springer VS), 277–280.

Organisation for Economic Co-operation and Development (2013). *PISA 2012 Ergebnisse. Was Schülerinnen und Schüler wissen und können: Schülerleistungen in Lesekompetenz, Mathematik und Naturwissenschaften, Bd. 1*. (Bielefeld: Bertelsmann).

Paetsch, J. (2016). *Der Zusammenhang zwischen sprachlichen und mathematischen Kompetenzen bei Kindern deutscher und bei Kindern nicht-deutscher Familiensprache* (Diss.). Freie Universität Berlin. Available online at: https://refubium.fu-berlin.de/handle/fub188/3610 (Accessed September 2, 2025).

Paetsch, J., Felbrich, A., and Stanat, P. (2015). Der Zusammenhang von sprachlichen und mathematischen Kompetenzen bei Kindern mit Deutsch als Zweitsprache. *Zeitschrift für Pädagogische Psychologie* 29, 19–29. doi: 10.1024/1010-0652/a000142

Paetsch, J., Radmann, S., Felbrich, A., Lehmann, R., and Stanat, P. (2016). Sprachkompetenz als Prädiktor mathematischer Kompetenzentwicklung von Kindern deutscher und nicht-deutscher Familiensprache. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie* 48, 27–41. doi: 10.1026/0049-8637/a000142

Peng, P., and Lin, X. (2019). The relation between mathematics vocabulary and mathematics performance among fourth graders. *Learn. Individ. Differ.* 69, 11–21. doi: 10.1016/j.lindif.2018.11.006

Peng, P., Lin, X., Ünal, Z. E., Lee, K., Namkung, J., Chow, J., and Sales, A. (2020). Examining the mutual relations between language and mathematics: a meta-analysis. *Psychol. Bull.* 146, 595–634. doi: 10.1037/bul0000231

Plath, J., and Leiss, D. (2018). The impact of linguistic complexity on the solution of mathematical modelling tasks. *ZDM* 50, 159–171. doi: 10.1007/s11858-017-0897-x

Prediger, S., Wilhelm, N., Büchter, A., Gürsoy, E., and Benholz, C. (2015). Sprachkompetenz und Mathematikleistung – Empirische Untersuchung sprachlich bedingter Hürden in den Zentralen Prüfungen. *Journal für Mathematik-Didaktik* 36, 77–104. doi: 10.1007/s13138-015-0074-0

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

Purpura, D. J., and Reid, E. E. (2016). Mathematics and language: Individual and group differences in mathematical language skills in young children. *Early Child. Res. Q.* 36, 259–268. doi: 10.1016/j.ecresq.2015.12.020

Raatz, U., and Klein-Braley, C. (1982). "The C-test – a modification of the cloze procedure," in *Practice and Problems in Language Testing IV*, eds. T. Culhane, C. Klein-Braley, and K. Douglas. Peterborough: Stevenson University of Essex, Dept. of Language and Linguistics.

Shohamy, E. (1982). Predicting speaking proficiency from Cloze tests: theoretical and practical considerations for tests substitution. *Appl. Linguist.* 3, 161–171. doi: 10.1093/applin/3.2.161

Sigott, G., and Köberl, J. (1996). "Deletion Patterns and C-Test Difficulty across Languages," in *Der C-Test. Theoretische Grundlagen und praktische Anwendungen, Bd. 2*, ed. G. Rüdiger (Bochum: Brockmeyer), 159–172.

Tabatabaei, O., and Shakerin, S. (2013). The effect of content familiarity and gender on EFL learners' performance on MC cloze test and C-test. *Int. J. Engl. Lang. Educ.* 1, 151–171. doi: 10.5296/ijele.v1i3.3952

Toll, S. W. M., and Van Luit, J. E. H. (2014). The developmental relationship between language and low early numeracy skills throughout kindergarten. *Except. Child.* 81, 64–78. doi: 10.1177/0014402914532233

Ufer, S., and Bochnik, K. (2020). The role of general and subject specific language skills when learning mathematics in elementary school. *Journal für Mathematik-Didaktik* 41, 81–117. doi: 10.1007/s13138-020-00160-5

Ufer, S., Leiss, D., Stanat, P., and Gasteiger, H. (2020). Sprache und Mathematik – theoretische Analysen und empirische Ergebnisse zum Einfluss sprachlicher Fähigkeiten in mathematischen Lern- und Leistungssituationen. *Journal für Mathematik-Didaktik* 41, 1–9. doi: 10.1007/s13138-020-00164-1

Ufer, S., Reiss, K., and Mehringer, V. (2013). "Sprachstand, soziale Herkunft und Bilingualität: Effekte auf Facetten mathematischer Kompetenz," in *Sprache im Fach: Sprachlichkeit und fachliches Lernen,* eds. M. Becker-Mrotzek, K. Schramm, E. Thürmann, and H. J. Vollmer (Münster: Waxmann), 185–201.

Viesel-Nordmeyer, N., Ritterfeld, U., and Bos, W. (2020). Welche Entwicklungszusammenhänge zwischen Sprache, Mathematik und Arbeitsgedächtnis modulieren den Einfluss sprachlicher Kompetenzen auf mathematisches Lernen im (Vor-)Schulalter? *Journal für Mathematik-Didaktik* 41, 125–155. doi: 10.1007/s13138-020-00165-0

Viljaranta, J., Lerkkanen, M.-K., Poikkeus, A.-M., Aunola, K., and Nurmi, J.-E. (2009). Cross-legged relations between task motivation and performance in arithmetic and literacy in kindergarten. *Learn. Instruct.* 19, 335–344. doi: 10.1016/j.learninstruc.2008.06.011

Wockenfuß, V. (2009). *Diagnostik von Sprache und Intelligenz bei Jugendlichen und jungen Erwachsenen*. Aachen: Shaker.

Wockenfuß, V., and Raatz, U. (2014). "Zur Validität von muttersprachlichen C-Tests: Bedeutung von verbaler Intelligenz und Informationsverarbeitungsgeschwindigkeit unter Berücksichtigung des Lebensalters," in *Der C-Test: Aktuelle Tendenzen*, ed. R. Grotjahn (Frankfurt am Main: Peter Lang), 119–224.