Do Students With Specific Learning Disorders With Impairments in Reading Benefit From Linguistic Simplification of Test Items in Science?

Nadine Cruz Neri¹ and Jan Retelsdorf¹

Abstract
Previous research illustrated that reading comprehension and science performance correlate highly. Because students with specific learning disorders with impairments in reading (SLD-IR) show deficits in reading comprehension, they may struggle to perform in science. As language in science is characterized by linguistic complexity, the question arises whether students with SLD-IR can be supported by reducing linguistic complexity. The aim of this preregistered study was to investigate whether students with SLD-IR benefit more from linguistic simplification in science than their peers without SLD-IR. The sample consisted of 70 students (age, M = 12.67; 50% female) with n = 35 having SLD-IR. Applying a multilevel logistic regression model, we found neither a main effect of linguistic simplification nor an interaction effect (differential boost) on science performance. However, students with SLD-IR performed significantly lower in science. Implications include further investigation on how to support students with SLD-IR in their science performance.

Learning science is essential in order to become an informed citizen and to make decisions regarding one’s health and social issues, such as global warming (Trefil, 2007). Because science education is often based on written text and test items (e.g., Norris & Phillips, 2003) due to its highly communicative nature (e.g., Yore et al., 2003), reading comprehension plays an important role for science performance. Considering the strong link between science performance and reading comprehension (e.g., Cromley, 2009), students with reading difficulties may particularly be affected because their science performance may be underestimated due to their reading deficits. In this article, we are particularly interested in the performance of students with specific learning disorders (SLD) with impairments in reading (SLD-IR) in a science test with test items that vary in their level of linguistic complexity, because studies focusing on the performance of students with SLD-IR in science are missing.

To lend students with SLD-IR support in overcoming their deficits, some test accommodations have been investigated, such as providing extra time (e.g., B. Lovett, 2010). However, to the best of our knowledge, there are no studies to date that focus on reducing the linguistic complexity of science test items that may interfere with the ability of students with SLD-IR to demonstrate their content knowledge. Such a test accommodation is referred to as linguistic simplification.

¹University of Hamburg

Corresponding Author: Nadine Cruz Neri, University of Hamburg, Von-Melle-Park 8, D-20146 Hamburg, Germany.
Email: nadine.cruz.neri@uni-hamburg.de
(LS; e.g., Rivera & Stansfield, 2004). It has been repeatedly investigated in the general education context with ambiguous results (e.g., Prophet & Badede, 2009). However, drawing on another special group of students, meta-analyses provide evidence that LS may be effective for weak readers, particularly for non-native students (e.g., Pennock-Roman & Rivera, 2011).

Given the importance of science education and the prevalence of students with SLD-IR (e.g., Landerl & Moll, 2010) who may have difficulties handling the linguistic demands of scientific texts and test items, the question remains if students with SLD-IR may particularly benefit from LS. To address this lacuna, we aim to investigate whether students with SLD-IR benefit more from LS in science than their peers without SLD-IR, resulting in a differential boost. Therefore, we conducted an experimental study for which we have developed science items in two linguistic versions: an original version that is inspired by academic and school science language (e.g., Fang, 2006) and an item version with (systematic) LS. To simplify the science items effectively, we drew on simplification studies (e.g., Bird & Welford, 1995), secondary analyses (e.g., Heppt et al., 2015), and guidelines (e.g., Kopriva, 2000).

The Importance of Reading Comprehension for Science Performance

Reading comprehension involves many different processes at different reading levels, starting with decoding individual letters and integrating them into a word through decoding the word meaning at the word level up to integrating these words into a coherent sentence at the sentence level or even into a coherent text passage at the text level (e.g., Kintsch, 1998). Research has extensively explored the importance of reading comprehension for overall academic success (e.g., Cooper et al., 2014), linking it to life outcomes, such as higher qualifications and socioeconomic status (e.g., Kutner et al., 2007).

Reading comprehension has further been shown to be strongly linked to science performance (e.g., Dempster & Reddy, 2007) due to the highly communicative nature of science learning (Yore et al., 2003). Reading scientific texts is very common in science education (e.g., Norris & Phillips, 2003), and hence, the importance of reading comprehension for science performance is apparent. This link has been extensively explored in research (e.g., Ozuru et al., 2009). The correlations within the school context range from .58 to .90 (Cromley, 2009; O’Reilly & McNamara, 2007), depicting high correlations of both skills (Cohen, 1992). However, not only reading comprehension at higher reading levels, that is, the text level, is important in order to perform well in science but also vocabulary at the word level because students need to deal with specific scientific vocabulary (e.g., Fang, 2006).

Students With SLD-IR

Given the importance of reading comprehension for science performance, it appears essential to support students in their reading skills acquisition for them to be able to perform well in science. This raises the question of how students with particularly poor reading skills perform in science measures. In the literature, several terms for students with poor reading skills have been established: poor, struggling, or low-achievement readers; students with a reading or learning disorder or dyslexia; students with reading or learning difficulties, and so on. Even though there are theoretical distinctions between some of these terms, it is somewhat difficult to differentiate between them in practice. Hence, these terms overlap and are often used synonymously (e.g., Coltheart & Prior, 2006).

For this study, we focus particularly on students with SLD-IR. The definitions of SLD-IR vary greatly (Tunmer & Greaney, 2010). For our study, we take the Diagnostic and Statistical Manual of Mental Disorders (5th ed.; DSM-5; American Psychiatric Association, 2013) as a basis for the definition of SLD-IR. The DSM-5 defines SLD-IR as a learning disorder that is characterized by impairments in reading despite having an (above) average
intelligence (American Psychiatric Association, 2013), which corresponds with the definition in the International Statistical Classification of Diseases and Related Health Problems by the World Health Organization (Dilling et al., 2015). However, some authors criticize these discrepancy criteria, arguing that these may depict artificial cut-off scores (Branum-Martin et al., 2013) or that intelligence measures often assess verbal intelligence, usually measured by reading tasks, which may in turn potentially bias overall intelligence scores (Pennington et al., 2019).

Students with SLD-IR are a very heterogeneous group that shows several deficits in a variety of subskills (e.g., Hock et al., 2009), including reading speed (e.g., Davies et al., 2013) and reading comprehension (e.g., Cárnio et al., 2017). Rose (2009) provides a detailed description of SLD-IR and the deficits of students with SLD-IR. According to Landerl and Moll (2010), worldwide prevalence rates vary between 4% and 9%, depicting the substantial number of affected students.

Considering the inseparable linkage between reading comprehension and science performance (e.g., Cromley, 2009), it seems plausible that students with SLD-IR struggle with science test items. Seifert and Espin (2012) suggest that such difficulties may arise from the mismatch between affected individuals’ reading skills and the reading requirements presupposed by science assessments. In view of the specialized language of school science (e.g., Fang, 2006), it is fair to assume that its linguistic complexity may hinder students with SLD-IR to answer test items correctly due to their difficulties in understanding the demands of these items. According to the cognitive load theory, which is challenged by some authors (see de Jong, 2010, for a critical discussion), the (potentially unfavorable) linguistic design and wording of test items may thus increase the cognitive load of the test items. Cognitive load refers to the amount of working memory resources that are needed to process an item (Sweller, 1988). If test items are presented unfavorably, for example, in terms of linguistic presentation, the test taker needs to use up additional valuable resources of working memory, which simultaneously decreases free capacities to process the technical content (Paas et al., 2010). Hence, students with SLD-IR may invest a lot of working memory capacity trying to decode the test item, leaving them with a rather limited capacity to actually process the technical content. Consequently, they may be hindered in showing their content knowledge due to deficits in reading comprehension (e.g., Cawthon et al., 2012). This circumstance is especially alarming when bearing in mind that most students with a learning disorder are expected to perform under the same conditions, or rather without accommodations for their deficits, as their peers without disabilities (e.g., Norman et al., 1998).

Linguistic Complexity in Science

Language and reading comprehension are integral parts of science performance (e.g., Yore et al., 2003). However, the language of science with its linguistic complexity differs from everyday language and tends to cause difficulties for the students performing in science tests (e.g., Fang, 2006). Several secondary analyses and experimental studies investigated linguistic complexity of science test items as well as the consequences on students’ performance (e.g., Cruz Neri et al., 2021). Several studies also illustrated that science items tend to incorporate unnecessary linguistic complexity at several levels that may decrease students’ performance, particularly for second-language learners (e.g., Martiniello, 2008).

Several linguistic features typical for science school language (e.g., Fang, 2006) might generate difficulties for students.

For instance, at the word level, the (technical) vocabulary (e.g., Fang, 2006) and the (low) frequency of words (e.g., Yap et al., 2009) are repeatedly discussed as potentially impeding comprehension and performance in science and related domains, such as mathematics. Pronouns (e.g., Fang, 2006), compounds (Heppt et al., 2015), and nominalizations (e.g., Dempster & Reddy, 2007) are further linguistic features considered to generate difficulties.
At the sentence level, features such as the sentence structure, including (complex) verb structures (e.g., Prophet & Badede, 2009), prepositions (e.g., Martiniello, 2008) and noun phrases (e.g., Fang, 2006), are discussed as generating difficulties for students and adults. For more details, see Text S1 in the supplementary material.

**LS as a Test Accommodation**

Taking into account the linguistic complexity that science items pose on students, it might be promising to provide students with SLD-IR with appropriate support in form of accommodations. Test accommodations are changes in testing to remove measurement errors that could arise from students’ disorders (e.g., Fuchs et al., 2000). Students with reading difficulties or SLD-IR may struggle with test items not due to their lack of content knowledge but rather due to the reading requirements of the test items (e.g., Cawthon et al., 2012). Hence, providing students with test accommodations attempts to reduce construct-irrelevant variance and measurement errors that would result from students’ disorders (e.g., Turkan & Liu, 2012). Well-designed test accommodations should therefore result in an interaction effect:

**Students with disorders should benefit more from accommodations than students without disorders, ensuring the validity of test measures for all students.**

This interaction effect is known as differential boost (e.g., Fuchs et al., 2000).

Prior researchers explored test accommodations for students with SLD-IR, such as providing extra time (e.g., B. Lovett, 2010). However, providing students only with accommodations such as extended time may not be sufficient for students to master the linguistic features of an item (e.g., Rhodes et al., 2015). Although a few studies focused on students with learning and intellectual disorders in general (e.g., Cawthon et al., 2012), to the best of our knowledge, no studies have yet been conducted that investigate test accommodations regarding LS particularly for students with SLD-IR. Hence, we need to draw on research related to other special groups of students.

**LS** refers to the modification of language in test items to reduce unnecessary linguistic complexity while preserving the technical content (e.g., Rivera & Stansfield, 2004). Many authors suggest simplifying the language used in test items for validity reasons (e.g., Wolf & Leon, 2009), and some even wrote practical guidelines for simplified language (e.g., Kopriva, 2000). Other authors, however, hold a critical view on LS. They argue “that specific linguistic knowledge is a component of content area mastery and thus that we cannot assume linguistic features are irrelevant to the target construct on content area tests without analyzing the use of the feature in the domain to which the test is intended to generalize” (Avenia-Tapper & Llosa, 2015, p. 108). We agree with this statement to some extent: Of course, LS has its limits. This is particularly the case when complex scientific principles are discussed because they need complex language to be described adequately. However, studies showed that scientific texts tend to use unnecessary linguistic complexity that can be reduced by means of LS without compromising the validity of an assessment (e.g., Abedi & Lord, 2001).

Simplification studies and guidelines usually suggest to modify several linguistic features simultaneously, because modifying one particular linguistic feature only is very difficult or even impossible. Therefore, LS usually includes the modification of several linguistic features at a time. Common strategies to achieve LS include replacing complex verb structures with an active voice and the present tense, omitting rare vocabulary, and simplifying sentence structures (see Abedi et al., 1997, for an overview).

Because it has been known that even slight linguistic modifications may affect students’ performance in science and similar subjects, such as mathematics (e.g., Cummins et al., 1988), many attempts of LS have been made. To date, there are LS studies for different domains, such as mathematics (e.g., Haag et al., 2015) and science (e.g., Prophet & Badede, 2009), but also for different populations, such as students with (learning) disorders (e.g., Kettler et al., 2012). Even though
the majority of these studies report higher students’ performances on linguistically simplified items, the effects do not always reach significance (e.g., Rivera & Stansfield, 2004). Meta-analyses investigating, among other accommodations, LS for English language learners report small but significant effects of simplifying language in test items (e.g., Pennock-Roman & Rivera, 2011).

In sum, it is evident that even minor changes in wording affect students’ performance significantly (e.g., Prophet & Badede, 2009), and although simplification studies have been conducted with students with specific learning and intellectual disorders (e.g., Cawthon et al., 2012; Fajardo et al., 2013), to the best of our knowledge, studies solely focusing on students with SLD-IR do not exist. To date, the studies investigating LS for students with disorders do not differentiate the effects of simplification between the different types of disorder (e.g., Elliott et al., 2010). We assume that it is of utter importance to differentiate between different types of disorders because students with intellectual disorders may face different problems and have different cognitive impairments compared with students suffering SLD-IR. We suggest that it is necessary to understand the linguistic features that generate difficulties for students with and without SLD-IR to support them better with their performance in science. Furthermore, understanding the mechanisms behind language characteristics in science items is also important for validity reasons, because high requirements in reading comprehension could hinder students with SLD-IR to show their content knowledge (e.g., Wolf & Leon, 2009). Following cognitive load theory, items should be designed in a way that keeps extraneous cognitive load, that is, the cognitive load caused by poorly designed items, to a minimum. In this way, cognitive resources can be used to master and process the test items rather than to deal with linguistically suboptimal item designs (Paas et al., 2010).

The Present Investigation

Building on previous research on LS of different populations (e.g., Abedi & Lord, 2001) and the importance of reading comprehension for science performance (e.g., Cromley, 2009), we aim to investigate whether students with SLD-IR benefit significantly from linguistically simplified items in science. Although many accommodations for students with SLD-IR have been investigated and many studies have investigated LS for students with disorders in general (e.g., Kettler et al., 2009) and for English language learners (e.g., Pennock-Roman & Rivera, 2011), no studies have particularly focused on accommodation by linguistically simplifying science items for students with SLD-IR. Again, we are aware that LS has its limits (see Avenia-Tapper & Llosa, 2015). However, we are convinced that LS may be able to reduce unnecessary linguistic complexity for students both with and without SLD-IR without compromising the validity of assessment. We assume that students with SLD-IR benefit significantly more from this type of accommodation because LS targets linguistic complexity and aims to facilitate reading. Thus, we aim to investigate whether there is a differential boost, that is, if LS boosts the science performance of students with SLD-IR significantly more than that of their peers without SLD-IR. For this purpose, we generated science items in an original version that was inspired by academic and school science language (e.g., Fang, 2006) and a simplified version.

To investigate our aims, we test the following hypotheses:

1. We hypothesize that overall, students without SLD-IR perform better on the science items than their peers with SLD-IR.
2. We expect that the total sample of students performs better on the simplified items than on the original version of science items.
3. We assume that although both students with and without SLD-IR benefit from the simplified science items, students with SLD-IR benefit significantly more from this type of accommodation but do not reach the level of students without SLD-IR.
Method

Sample

The data collection for the main study took place between February and November 2021. Participants were mainly recruited from schools of the nonacademic track in different federal states of Germany. The nonacademic track is considered the general education, which enables students to start a vocational apprenticeship (for more detailed description of the German schooling system, see Trautwein et al., 2006). We recruited additional participants with SLD-IR from institutions for special-needs language education (visited besides regular school). Students received €10 for their participation.

For the preregistration of the study, we conducted a power analysis for multilevel logistic regression (Olvera Astivia et al., 2019), using R (R Core Team, 2009–2019), before starting the data collection. Results showed that a total sample of \( n = 70 \) students at the between level was required to achieve a power of 0.90 or higher.

In total, \( N = 203 \) seventh graders participated in the study. However, 123 students needed to be excluded from the analytic sample because they did not meet our inclusion criteria. First, 15 students skipped at least one subscale completely. Second, 16 did not answer the attention-check questions correctly. Third, 90 students had a \( T \) value below 40 in the cognitive measures. Finally, four students’ mean effort in completing the study fell two standard deviations below the mean effort of all students. However, one of these students reached above-average performance in both cognitive abilities and reading comprehension and, hence, was included in the analytic sample.

This resulted in a sample of \( N = 79 \) students (35 without SLD-IR; 44 with SLD-IR). We defined SLD-IR to be present if students show impairments in their reading skills while not underperforming in cognitive measures, to ensure that potential impairments in reading are not the consequence of potential intellectual disorders (American Psychiatric Association, 2013). Based on our power analysis, we aimed for an equal distribution of students with and without SLD-IR. Hence, we randomly selected 35 students with SLD-IR for our statistical analyses. We also conducted the analyses with all \( N = 79 \) students. The results were cum grano salis the same: Solely the effect of gender on science performance was nonsignificant (\( B = -0.065, p = .629 \)).

This resulted in an analytic sample of \( n = 70 \) students (age, \( M = 12.67, SD = 0.58; 50\% \) female). Thereof, 50% had SLD-IR according to our definition. Most participants stated that they speak only German at home (91.40%). Only 8.60% of students stated that they speak German and another language or speak only another language at home. There were no missing data on the variables that we included in our statistical model.

The students with SLD-IR and students without SLD-IR did not differ regarding their age, \( t(68) = -0.20, p = .839 \); gender, \( \chi^2(1) = 0.51, p = .473 \); and migrant background, \( \chi^2(2) = 3.40, p = .183 \). Furthermore, they did not differ regarding their grades in the subjects German, \( t(62.88) = -0.98, p = .330 \), and science, \( t(65) = -0.07, p = .945 \). At this point, one must note that a considerable portion of students did not state their grades in those subjects (German, 18.60%; science, 28.30%).

Measures

The study consisted of four measurements: reading comprehension, cognitive abilities, science items, and a short demographic questionnaire. We chose test methods with a single-choice format for every measure because we wanted to focus on receptive reading skills. Open-response formats measure productive skills besides receptive reading skills, usually decreasing performance (e.g., Härtig et al., 2015).

The main study was conducted online (see Text S2 in the supplementary material). As applied in the Programme for International Students Assessments (Kunter et al., 2002), we implemented an item after every test measure asking the students to state how much effort they made on a scale from 1 to 10 (see “Effort Control” at https://osf.io/
Because we conducted the study in Germany, all measures are available only in German language.

**Reading Comprehension.** Participants completed the German version of the Reading Comprehension Test for Grade 1 to 7—Version II (in German, Ein Leseverständnistest für Erst- bis Siebtklässler—Version II; Lenhard et al., 2018), which assesses students’ reading comprehension, fluency, and accuracy. The assessment is divided into three subtests assessing the mentioned measurements at the word, sentence, and text levels. Students had 2 to 6 min to complete each subtest containing 26 to 75 items. For students to be presumed to suffer from SLD-IR, they need to perform at a T value of or less than 35 (Lenhard et al., 2018).

**Cognitive Abilities.** Students’ cognitive abilities were assessed with a nonverbal subtest of a revised German test (in German, Kognitiver Fähigkeitstest 4-12 + R; Heller & Perleth, 2000), which is considered to be a fair indicator of general cognitive abilities (Neisser et al., 1996). The test is based on the Cognitive Abilities Test (Thorndike & Hagen, 1971). The subtest consists of 25 items. Students had 8 min to work on the subtest. Students performing at a T value of less than 40 were considered as underperforming in cognitive measures.

**Science Items.** The items were generated according to the school curricula of the nonacademic track of different German federal states. To fit the structure of German school curricula, the science items test the subject matter of Grades 5 and 6 that was consistent across the school curricula of the nonacademic track of different federal states. The science items were tested in a pilot study beforehand (see Text S3 and Figure S1 in the supplementary material).

The science items were systematically and linguistically modified, resulting in two linguistic versions of each item with the same content knowledge: the original version and a simplified version. The items of both versions can be seen in the document “Science Items in Both Versions” at https://osf.io/kta6c. To compare the original and the simplified items, we coded several linguistic features with the software LATIC (Linguistic Analyzer for Text and Item Characteristics; Cruz Neri & Klückmann, 2021) that were considered when developing the items. The descriptives of both science versions are provided in Table S1 and discussed in Text S4 in the supplementary material. At https://osf.io/kta6c, a detailed description and an example of coding are provided in the files “Coding of Linguistic Features” and “Example of Coded Item.”

Students worked on 20 science items, including 10 original and 10 simplified items, respectively.

**Both versions.** The science items of both versions have multiple characteristics in common regarding layout and format. We followed guidelines for generating the items (e.g., Haladyna et al., 2002) with the intent to keep characteristics regarding the general structure alike between both versions. The technical content was the same, which was double-checked by a science teacher and a science teacher student. The 20 items include the topics of living beings, including humans, fauna, and flora (n = 8); sensory organs (n = 5); weather and climate (n = 3); and the solar system (n = 4), which were explicitly stated as general educational goals in the curricula of Grades 5 and 6. The students had 25 min to work on the items.

All items were presented in a single-choice format with four response choices to only measure the students’ receptive language skills rather than set additional demands on their productive language skills, which would be needed in open-response formats (e.g., Härtig et al., 2015). The items were solely based on words; no additional pictures were generated for the items due to possible confounding effects on reading comprehension and, consequently, science performance (e.g., Carney & Levin, 2002). Further criteria that we followed when generating the items in both versions are presented in Table S2 of the supplementary material.

**Original version.** We followed no guidelines or recommendations regarding linguistic features to make items more accessible for students but rather took up characteristics of
academic and school science language (e.g., Fang, 2006). The original version was developed with the intent to resemble science items that could be used in science classes and large assessment studies.

Simplified version. The original items were systematically simplified following recommendations of several guidelines (e.g., Kopriva, 2000) and taking into account various results of different studies and secondary analyses (e.g., Haag et al., 2013), but the technical content remained the same. For the simplification of science items and their response choices, we established the criteria stated in Table 1.

Randomization. All students worked on the four measures: reading comprehension, the subtest of the cognitive abilities measure, the science items, and the short questionnaire. The order of the first three measures was randomized. All students completed the short questionnaire last.

For the study, we prepared two different science test versions containing the same technical content in each test version. We made sure that the word count in both test versions was similar (version A, 923 words; version B, 930 words). To further ensure equivalence between both test versions, we allocated science items taking their difficulty based on the pilot-testing into account (see Text S3 and Figure S1 in the supplementary material). Drawing on the probabilities of correct responses in the pilot study, test versions A ($M = .31$, $SD = .12$) and B ($M = .32$, $SD = .13$) were comparable. The order of science items was randomized.

Ethical Approval

The study design was approved by the responsible ministries and authorities of the federal states we collected our data in. The study was approved by the ethics committee of the University of Hamburg. It was conducted in compliance with the ethical guidelines of the American Psychological Association, and all subjects gave written informed consent.

| Table 1. Guidelines for Generating the Linguistically Simplified Version of Science Items. |
|---|
| **Word level** |
| 1. Words should be as simple and high frequent as possible. |
| 2. Pronouns should be used as little as possible. If pronouns are used, it is made clear which noun they are referring to. |
| 3. Compounds are omitted if they are not scientific technical terms. |
| 4. Nominalizations should be avoided unless it is a (sub)technical or general term. |
| 5. Synonyms should be eliminated. Rather, the same words should be used to refer to the same concept, phenomenon, technical term, and so on. |
| **Sentence level** |
| 1. The sentence structure should be kept simple, containing mainly main clauses. Complex sentence structures, such as sentences with embedded clauses or parentheses, should be omitted completely. |
| 2. The average sentence length should be rather short. |
| 3. Items should be phrased in an active voice and in the present tense. |
| 4. The number of prepositions should be kept to a minimum. |
| 5. Expanded noun phrases should be avoided or at least kept as short as possible. |
| 6. Negations should be kept to a minimum. In case negations cannot be removed, the negation should be capitalized and boldfaced. |
| 7. The genitive should be omitted. Dative phrasing is to be kept to a minimum. |
| **Text level** |
| 1. The word count in an item should be kept to a minimum, eliminating information that is irrelevant for solving the item correctly. Short items should be better for students with SLD-IR due to their deficits in reading competencies. |

Statistical Analyses

To test our hypotheses, we analyzed the interaction effect of students’ (nonexisting) SLD-IR and the science items’ version on students’ performance by applying a multilevel logistic regression model using the software Mplus 8.7 (Muthén & Muthén, 1998–2017). We conducted a multilevel regression model due to the hierarchy of our data: We consider science items as being nested in students. Moreover,
we apply logit regressions due to the binary nature of our dependent variable, which is the science performance on item level (0 = false response, 1 = correct response). In our model, the item version (original items vs. simplified version) served as a within-subjects variable, and students’ (nonexisting) SLD-IR served as a between-subjects variable (0 = student without SLD-IR, 1 = student with SLD-IR).

To test the interaction effect of both these variables, we included a cross-level interaction on science performance. We further controlled for two dummy-coded variables: gender (0 = male, 1 = female) and migratory background (0 = native speaker, 1 = non-native speaker). The odds ratios were calculated.

## Results

### Descriptive Statistics, Reliabilities, and Correlations

Overall, students reported relatively high efforts in completing the study ($M = 7.89$, $SD = 1.80$). The descriptive statistics and Cronbach’s alphas of students’ reading comprehension, cognitive abilities, and science performance are depicted in Table 2. Table 3 further includes the correlations of both assessment and demographic variables.

The probability of a correct response was $M = .62$ ($SD = .18$) for the original science items and $M = .67$ ($SD = .21$) for the simplified science items. This corresponds to a small effect ($d = 0.26$), according to Cohen (1988). The differences between the probability of a correct response ranged from $-0.20$ to $0.40$ ($M = .05$, $SD = .15$), a negative difference meaning that the original item was solved correctly more often. The probabilities of a correct response of the science items are depicted in Figure 1.

## Multilevel Analysis

The results of the multilevel analysis are presented in Table 4. Corroborating our first hypothesis, students with SLD-IR performed significantly lower than their peers without SLD-IR. Their chance of solving a science item correctly was 34.4% lower compared with their peers without SLD-IR. Our second and third hypotheses, however, were not corroborated. The overall sample did not perform better on the simplified version of items (main effect), nor did students with SLD-IR benefit significantly more from the simplified version than students without SLD-IR (interaction effect).

## Discussion

The aim of this study was to investigate whether the LS of science items provides adequate support for students with SLD-IR, resulting in a differential boost. For this, we generated two

| Variable               | Descriptive statistics |          |          |          | Maximum score | Cronbach’s alpha |
|------------------------|------------------------|----------|----------|----------|---------------|-----------------|
|                       | Students with SLD-IR   | Students without SLD-IR | Total |          |               |                 |
|                       | $M$ | $SD$ | $M$ | $SD$ | $M$ | $SD$ |               |                 |
| Reading comprehension  |          |          |          |          |               |                 |
| Word level             | 38.49 | 10.48 | 52.37 | 9.71 | 45.43 | 12.23 | 75.00 | .962 |
| Sentence level         | 17.89 | 4.48 | 25.40 | 3.96 | 21.64 | 5.65 | 36.00 | .915 |
| Text level             | 15.89 | 3.97 | 21.91 | 2.58 | 18.90 | 4.50 | 26.00 | .867 |
| Total                  | 72.26 | 14.23 | 99.69 | 13.12 | 85.97 | 19.38 | 137.00 | .968 |
| Cognitive abilities    | 17.63 | 4.99 | 17.09 | 4.56 | 17.34 | 4.75 | 25.00 | .811 |
| Science                | 11.94 | 2.84 | 13.29 | 3.15 | 12.61 | 3.05 | 20.00 |               |

Note. SLD-IR = specific learning disorders with impairments in reading.
versions of science items: science items inspired by academic and school science language (e.g., Fang, 2006) and (linguistically) simplified items. Although our results showed a higher probability of a correct response in the simplified science items, these results did not reach significance. A differential boost was also not found.

The Effect of Students’ SLD-IR on Science Performance

The results illustrated that students without SLD-IR significantly outperformed their peers with SLD-IR in the science assessment, corroborating our first hypothesis. Other students’ characteristics, such as cognitive abilities, do not seem to diminish this effect of students’ SLD-IR status (we tested a model including cognitive abilities as a further covariate; see Table S5 of the supplementary material). As noted earlier, this result seems plausible given that science is a highly communicative domain (Yore et al., 2003) and items are usually presented in written text. In our study, we presented students with science items that were heavily text based.

Table 3. Correlations Between the Variables in the Main Study.

| Variable               | 1.1  | 1.2  | 1.3  | 1.4  | 2   | 3   | 4   | 5   |
|------------------------|------|------|------|------|-----|-----|-----|-----|
| 1. Reading comprehension | —    | —    | —    | —    | —   | —   | —   | —   |
| 1. Word level          | —    | —    | —    | —    | —   | —   | —   | —   |
| 1. Sentence level      | .715** | —    | —    | —    | —   | —   | —   | —   |
| 1. Text level          | .328** | .764** | —    | —    | —   | —   | —   | —   |
| 1. Total               | .916** | .921** | .662** | —    | —   | —   | —   | —   |
| 2. Cognitive abilities | —    | —    | —    | —    | —   | —   | —   | —   |
| 3. Science             | —    | —    | —    | —    | —   | —   | —   | —   |
| 4. SLD-IR              | —    | —    | —    | —    | —   | —   | —   | —   |
| 5. Gender              | —    | —    | —    | —    | —   | —   | —   | —   |
| 6. Migratory background| —    | —    | —    | —    | —   | —   | —   | —   |

Note. SLD-IR (0 = student without SLD-IR, 1 = student with SLD-IR), gender (0 = male, 1 = female), and migratory background (0 = native speaker, 1 = non-native speaker) were dummy coded. SLD-IR = specific learning disorders with impairments in reading.

*p < .05. **p < .01 (two tailed).

Figure 1. Probabilities of correct responses of the original and the simplified items (main study). Note: The continuous line depicts the averaged probability of a correct response of both item versions.
without illustrations that might help students understanding the technical content of the items. Hence, proficient reading was needed to solve the science items. Students with SLD-IR might have performed worse than their peers without SLD-IR due to a mismatch between their reading comprehension and the reading requirements presupposed by the science items (see Seifert & Espin, 2012). Drawing on cognitive load theory (Sweller, 1988), one might infer that students with SLD-IR needed more cognitive resources to read and process the science items compared with their peers without SLD-IR due to their impairments in reading. As proposed by the construction-integration model (e.g., Kintsch, 1998), this seems plausible: As reading processes on the surface structure (i.e., the linguistic presentation in terms of linguistic complexity) consume a lot of resources for students with SLD-IR, only a limited capacity remains to build an adequate situational model. As a result, students with SLD-IR might have only a limited capacity left to process the technical content of the item.

The assumption that students with SLD-IR have a higher cognitive load due to the reading demands of the science items should be investigated in future research. Although it is still unclear how cognitive load is best measurable (e.g., DeLeeuw & Mayer, 2008), a subjective measure of cognitive load from students’ perspective could bring more insight into where exactly difficulties in solving science items arise for students with SLD-IR. Drawing on Paas (1992), students with SLD-IR could be directly asked how much mental effort they needed to invest (a) to read and comprehend the presented item (extraneous load) as well as (b) to process the technical content of the science items (intrinsic load).

### The Missing Effect of LS

Whereas students’ status of SLD-IR affected their science performance significantly, the LS did not. Although all students performed slightly better on the simplified science items, contrary to our expectations, this difference was not significant. The LS also did not result in a differential boost, again, not corroborating our hypothesis. Thus, although it seems as if students with SLD-IR needed additional cognitive resources to process the science items compared with their peers without SLD-IR, the LS did not reduce the cognitive load of science items (sufficiently) to support the students with SLD-IR significantly.

Looking into the probabilities of a correct response, some items were easier to solve in the original version. This is somewhat surprising given the results of our pilot study (Figure S1 in the supplementary material). One must note, however, that the pilot sample had an overall lower probability of correct responses on the science items than the analytic sample of the main study did: Their probability of correct responses was less than half the probability of correct responses of their peers in the main study. It cannot be ruled out that the different science proficiency levels might explain the differences in the probabilities of correct responses of the pilot and main study samples.

On the one hand, our finding that LS might not be significantly helpful for a general population is in line with prior research (e.g., Bird

### Table 4. Results of the Multilevel Analyses Predicting Science Performance (Unstandardized Parameters).

| Predictor               | B     | SE   | p     | OR    |
|-------------------------|-------|------|-------|-------|
| Intercept               | -0.967| 0.177| <.001 |       |
| Level 1 Science item    | 0.211 | 0.183| .249  | 1.235 |
| version                 |       |      |       |       |
| Level 2 SLD-IR          | -0.421| 0.188| .025  | 0.656 |
| Gender                  | -0.459| 0.161| .004  | 0.632 |
| Migratory background    | -0.275| 0.362| .448  | 0.760 |
| Interaction effect      |       |      |       |       |
| Science Item Version ×  | 0.139 | 0.234| .554  | 1.149 |
| SLD-IR                  |       |      |       |       |

Note. Science item version (0 = original, 1 = simplified), SLD-IR (0 = student without SLD-IR, 1 = student with SLD-IR), gender (0 = male, 1 = female), and migratory background (0 = native speaker, 1 = non-native speaker) were dummy coded. OR = odds ratio; SLD-IR = specific learning disorders with impairments in reading.
On the other hand, the results not corroborating our hypothesis of a differential boost are more difficult to interpret. Because our study is—to the best of our knowledge—the first one to investigate LS particularly for students with SLD-IR, we can only propose assumptions as to why LS failed to significantly provide support for students with SLD-IR.

First, students with SLD-IR might struggle with the science items in both versions due to their deficits in reading. Potentially the LS was not effective for these students because they still had difficulties in processing the science items in the simplified version due to their significant impairment in reading; LS, which targets linguistic complexity, might not be a sufficient accommodation to compensate for students’ SLD-IR. However, students with SLD-IR, despite having the same diagnosis, are a heterogeneous group (Hock et al., 2009). Hence, LS might be helpful for students with SLD-IR who show deficits in their reading comprehension or vocabulary. Students with SLD-IR having primarily deficits in their reading speed or fluency, however, might not benefit from LS because this accommodation does not target the roots of their impairment in reading. Thus, it might not be sufficient to investigate LS for all students with SLD-IR rather than for specific subgroups of students.

Second, when simplifying the science items, we looked into empirical results for a general student population and specific subgroups, such as non-native speakers (e.g., Haag et al., 2013). We also followed several guidelines (e.g., Kopriva, 2000). However, there is little empirical evidence to corroborate the recommendations stated in these guidelines. Hence, there is a chance that we modified linguistic features that might not even generate difficulties for students with SLD-IR in the first place. For instance, although there is evidence that noun phrases generate difficulties in comprehension and performance for non-native speakers (e.g., Martiniello, 2008), there is no evidence that native speakers or other groups of students face the same difficulties. Nevertheless, we chose to consider noun phrases in our simplification approach due to the lack of research regarding students with SLD-IR.

In simplifying the items the way we did, some simplified science items might differ from the items the students usually are confronted with in school, and hence, they might not be used to the approaches we followed. In this respect, one item stands out: It was solved 20% more often in the original than in the simplified version (Item 8, \(M_{\text{original}} = .62, M_{\text{simplified}} = .42\)). In hindsight, we noticed that the original version was heavily contextualized, which is common in science testing (Ruiz-Primo & Li, 2016). Contextualized items provide students with a context facilitating solving a task due to the items being more realistic and less abstract (Haladyna, 1997). The fact that the eighth science item was contextualized in the original version, but not in the simplified version, might explain why students were better able to solve the item in the original version. Thus, it is necessary for future research not only to identify linguistic features that might facilitate or impede the performance of students with SLD-IR but also to identify further item characteristics that go beyond the linguistic presentation.

Third, prior research provides evidence that LS in science items might be helpful for specific subgroups, such as students with academic difficulties (e.g., Elliott et al., 2010) and non-native speakers (e.g., Bird & Welford, 1995). However, students with SLD-IR might need additional support alongside LS. In a study by Kettler et al. (2012) investigating LS for students with and without disorders, all students benefited from LS and performed better on the simplified science items. However, the researchers not only modified science items in terms of linguistic complexity but also, for instance, added illustrations. Hence, it might be possible that LS alone is not sufficient to provide adequate support for students with SLD-IR in science. In providing illustrations for students with disorders, Kettler et al. made use of the multimedia effect (Mayer, 2005). This effect refers to the phenomenon...
that more information can be acquired when learning information is presented both verbally and visually because these two types of information are processed independently. Studies indicate that the combination of both LS and the multimedia effect positively affects science performance for a general student population (e.g., Siegel, 2007), non-native students (e.g., Noble et al., 2020), and students with learning disorders (Kettler et al., 2012). Hence, combining LS and the multimedia effect to reduce linguistic complexity in support of students with SLD-IR might be a promising project to pursue in future research.

**Theoretical and Practical Implications**

Although our hypotheses regarding the effect of LS were not corroborated, there are still theoretical and practical implications than can be derived. We want to particularly focus on implications (a) for future research on LS and (b) for adequately supporting students with SLD-IR in science education.

As to future research, we consider it important to further investigate LS. At this point, we still need to examine (a) whether LS is not an adequate accommodation for students with SLD-IR at all or (b) whether LS works only under specific circumstances.

*We still assume that LS as an accommodation has potential to adequately (and significantly) support students with SLD-IR.*

Thus, the nonsignificant results should not discourage other researchers to pursue the investigation of LS for students with SLD-IR, because our study was the first attempt to investigate LS for this specific student population. As noted earlier, we propose three possible directions for future research in this regard.

First, LS might be especially helpful for a specific subgroup of students with SLD-IR, such as students with specific impairments in their reading comprehension. Second, other item characteristics that might facilitate or impede science performance need to be examined. For instance, contextualizing science items is assumed to facilitate science performance (Haladyna, 1997). In our study, we focused only on simplifying science items linguistically but did not take other item characteristics, such as the contextualization, into account, which might have affected the effects of LS. Finally, we would encourage research investigating the combination of LS with other ways of supporting students. Especially the combination of LS and the multimedia effect (Mayer, 2005) seems to be a promising method, with some studies providing first evidence for students with learning disorders (e.g., Kettler et al., 2012).

As a second implication, it is important to find ways of supporting students with SLD-IR in their science performance because their status of SLD-IR affected science performance significantly. Although little research has focused on students with SLD-IR in science education, much more is known about students with all types of learning disorders. In two meta-analyses, it has been illustrated that the science performance of students with learning disorders can significantly improve, first, when they are provided with supplemental instructions, such as mnemonics (Therrien et al., 2011), and, second, when being taught the science curriculum through the usage of graphic organizers (Dexter et al., 2011). It is further known that students with learning disorders benefit from structured and direct instruction by their teachers in science education (Therrien et al., 2011). For instance, Seifert and Espin (2012) found that students with SLD significantly benefited from word recognition activities guided by a teacher prior to reading a science text; their vocabulary learning and reading fluency increased significantly. Thus, there seems to be promising accommodations for the overall group of students with SLD. However, further research is needed to investigate whether these instructional strategies also support students with SLD-IR in science.

Ultimately, it would be desirable to support students with SLD-IR in improving their reading skills in general to address their deficits at their root. In a recent study, M. Lovett et al.
(2021) examined a long-term, multicomponent reading intervention for students in Grades 6 to 8: Students with SLD-IR participating in the intervention significantly outperformed their peers in the control group (receiving reading instructions by special education teachers) in multiple measures, such as text comprehension. Considering the high correlations of reading and science performance (e.g., Cromley, 2009), it can be assumed that the positive effects on reading comprehension might also affect science performance positively.

**Limitations**

Despite having considerable strengths, this study has limitations that need to be addressed. First, due to the ongoing COVID-19 pandemic at the time, the study was conducted online rather than—as originally planned—in face-to-face settings. Therefore it could not be ensured that participants worked on the assessments by themselves and did not receive any help by significant others. This might be especially true for the science items, given that we repeatedly got the feedback that students finished the items way before the time limit of 25 min was over. Hence, we cannot guarantee that students did not look for answers online or asked others for help. As mentioned in Text S2 in the supplementary material, multiple approaches were taken in order to mitigate the risk of students not working on their own (e.g., Olt, 2002). Still, there is no guarantee that students were unassisted.

Second, we cannot rule out that the main effect of SLD-IR can be traced back to students’ ability to guess the correct answers. Scruggs and Lifson (1984) illustrated that students with learning disorders are significantly better at guessing at random than their peers without learning disorders. The authors assume that this effect might partly be explained by their lower reading comprehension and other factors, such as attention deficits or test anxiety.

Third, the students in the main study did quite well on the science items on average, especially compared with the students of the pilot study. Hence, the missing effect of LS might possibly be traced back to the main study students’ higher performance in science.

Finally, the missing (main and interaction) effects of LS on science performance might be due to our approach in simplifying the science items. As noted earlier, we noticed that in simplifying the science items, we sometimes removed the context, which might impede comprehension and performance instead of facilitating it. Furthermore, we needed to follow guidelines and empirical results of other student populations, such as non-native speakers (e.g., Kopriva, 2000), due to the lack of research regarding students with SLD-IR. Future research is needed to investigate which linguistic features and other item characteristics, such as the contextualization of science items, facilitate or impede science performance of students with SLD-IR.

**Conclusion**

The present study showed that students without SLD-IR significantly outperform their peers with SLD-IR in science assessments. The attempt to linguistically simplify the science items to support students with SLD-IR was not successful: We did not find a main effect of LS nor an interaction effect of students’ SLD-IR status and LS (differential boost) on science performance. Thus, LS does not seem to provide significant support for students with SLD-IR in their science performance. Still, we are confident that LS has potential, and hence, we encourage future research to investigate LS further. Identifying specific (linguistic) item characteristics that might boost or impede science performance and investigating the combination of LS with other approaches (such as the multimedia effect) might be promising strategies to support students with SLD-IR adequately in the future.

**References**

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219–234. https://doi.org/10.1207/S15324818AME1403_2
Abedi, J., Lord, C., & Plummer, J. R. (1997). *Final report of language background as a variable in NAEP mathematics performance* (CSE
Cruzer Neri and Retelsdorf

Technical Report No. 429. National Center for Research on Evaluation, Standards, and Student Testing. https://cresst.org/wp-content/uploads/TECH429.pdf

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders.

Avenia-Tapper, B., & Llosa, L. (2015). Construct relevant or irrelevant? The role of linguistic complexity in the assessment of English language learners’ science knowledge. Educational Assessment, 20(2), 95–111. https://doi.org/10.1080/10627197.2015.1028622

Bird, E., & Welford, G. (1995). The effect of language on the performance of second-language students in science examinations. International Journal of Science Education, 17(3), 389–397. https://doi.org/10.1080/0950069950170309

Branum-Martin, L., Fletcher, J. M., & Stuebing, K. K. (2013). Classification and identification of reading and math disabilities: The special case of comorbidity. Journal of Learning Disabilities, 46(6), 490–499. https://doi.org/10.1177/0022219412468767

Carney, R. N., & Levin, J. R. (2002). Pictorial illustrations still improve students’ learning from text. Educational Psychology Review, 14(1), 5–26. https://doi.org/10.1023/A:1013176039260

Cárnio, M. S., Vosgrau, J. S., & Soares, A. J. C. (2017). The role of phonological awareness in reading comprehension. Revista CEFAC, 19(5), 560–600. https://doi.org/10.1590/1982-0216201619518316

Cawthon, S. W., Kaye, A. D., Lockhart, L. L., & Beretvas, S. N. (2012). Effects of linguistic complexity and accommodations on estimates of ability for students with learning disabilities. Journal of School Psychology, 50(3), 293–316. https://doi.org/10.1016/j.jsp.2012.01.002

Cohen, J. (1988). Statistical power analysis for behavioral sciences (2nd ed.). Lawrence Erlbaum.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Coltheart, M., & Prior, M. (2006). Learning to read in Australia. Australian Journal of Learning Difficulties, 11(4), 157–164. https://doi.org/10.1080/19404150609546820

Cooper, B. R., Moore, J. E., Powers, C. J., Cleveland, M., & Greenberg, M. T. (2014). Patterns of early reading and social skills associated with academic success in elementary school. Early Education and Development, 25(8), 1248–1264. https://doi.org/10.1080/10409289.2014.932236

Cromley, J. (2009). Reading achievement and science proficiency: International comparisons from the Programme on International Student Assessment. Reading Psychology, 30(2), 89–118. https://doi.org/10.1080/02702710802274903

Cruz Neri, N., Guil, K., & Retelsdorff, J. (2021). Language in science performance: Do good readers perform better? European Journal of Psychology of Education, 36(1), 45–61. https://doi.org/10.1080/s10212-019-00453-5

Cruz Neri, N., & Klückmann, F. (2021). LATIC: A linguistic analyzer for text and item characteristics (Version 1.1.0) [Computer software]. https://github.com/florianklueckmann/LATIC

Cummins, D. D., Kintsch, W., Reusser, K., & Weinier, R. (1988). The role of understanding in solving word problems. Cognitive Psychology, 20(4), 405–438. https://doi.org/10.1016/0010-0285(88)90011-4

Davies, R., Rodriguez-Ferreiro, J., Suárez, P., & CUetos, F. (2013). Lexical and sub-lexical effects on accuracy, reaction time and response duration: Impaired and typical word and pseudoword reading in a transparent orthography. Reading and Writing, 26(5), 721–738. https://doi.org/10.1007/s11145-012-9388-1

de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. Instructional Science, 38(2), 105–134. https://doi.org/10.1007/s11251-009-9110-0

DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. Journal of Educational Psychology, 100(1), 223–234. https://doi.org/10.1037/0022-0663.100.1.223

Dempster, E. R., & Reddy, V. (2007). Item readability and science achievement in TIMSS 2003 in South Africa. Science Education, 91(6), 906–925. https://doi.org/10.1002/sce.20225

Dexter, D. D., Park, Y. J., & Hughes, C. A. (2011). A meta-analytic review of graphic organizers and science instruction for adolescents with learning disabilities: Implications for the intermediate and secondary science classroom. Learning Disabilities Research & Practice, 26(4), 204–213. https://doi.org/10.1111/j.1540-5826.2011.00341.x

Dilling, H., Mombour, W., & Schmidt, M. H. (2015). Internationale Klassifikation psychischer Störungen: ICD–10 Kapitel V (F) [International statistical classification of diseases and related health problems: ICD–10 chapter V (F)]. Hogrefe.
Elliott, S. N., Kettler, R. J., Beddow, P. A., Kurz, A., Compton, E., McGrath, D., Bruen, C., Hinton, K., Palmer, P., Rodriguez, M. C., Bolt, D., & Roach, A. T. (2010). Effects of using modified items to test students with persistent academic difficulties. *Exceptional Children, 76*(4), 475–495. https://doi.org/10.1177/001440291007600406

Fajardo, I., Tavares, G., Ávila, V., & Ferrer, A. (2013). Towards text simplification for poor readers with intellectual disability: When do connectives enhance text cohesion? *Research in Developmental Disabilities, 34*(4), 1267–1279. http://doi.org/10.1016/j.ridd.2013.01.006

Fang, Z. (2006). The language demands of science reading in middle school. *International Journal of Science Education, 28*(5), 491–520. https://doi.org/10.1080/09500690500339092

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C. L., Binkley, E., & Crouch, R. (2000). Using objective data sources to enhance teacher judgments about test accommodations. *Exceptional Children, 67*(1), 67–81. https://doi.org/10.1177/001440290006700105

Haag, N., Hepp, B., Roppelt, A., & Stanat, P. (2015). Linguistic simplification of mathematics items: Effects for language minority students in Germany. *European Journal of Psychology of Education, 30*(2), 145–167. https://doi.org/10.1007/s10212-014-0233-6

Haag, N., Hepp, B., Stanat, P., Kuhl, P., & Pant, H. A. (2013). Second language learners’ performance in mathematics: Disentangling the effects of academic language features. *Learning and Instruction, 23*(3), 24–34. https://doi.org/10.1016/j.learninstruc.2013.04.001

Haladyina, T. M. (1997). *Writing test items to evaluate higher order thinking*. Allyn and Bacon.

Haladyina, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5

Härtig, H., Heitmann, P., & Retelsdorf, J. (2015). *Analyse der Aufgaben zur Evaluation der Bildungsstandards in Physik: Differenzierung von schriftsprachlichen Fähigkeiten und Fachlichkeit* [Analyses of the tasks for evaluating the educational standards in physics: Differentiation between written language proficiency and content knowledge]. *Zeitschrift für Erziehungswissenschaft, 18*(4), 763–779. https://doi.org/10.1007/s11618-015-0646-2

Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeits test für 4. bis 12. Klassen*, Revision: *KFT 4-12+ R* [Technical manual]. Beltz Test.

Hepp, B., Haag, N., Böhme, K., & Stanat, P. (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

Hock, M. F., Brasseur, I. F., Deshler, D., Catts, H. W., Marquis, J. G., Mark, C. A., & Stribling, J. W. (2009). What is the reading component skill profile of adolescent struggling readers in urban schools? *Learning Disability Quarterly, 32*(1), 21–38. https://doi.org/10.2307/25476660

Kettler, R. J., Dickenson, T. S., Bennett, H. L., Morgan, G. B., Gilmore, J. A., Beddow, P. A., Swaffield, S., Turner, L., Herrera, B., Turner, C., & Palmer, P. W. (2012). Enhancing the accessibility of high school science tests: A multistate experiment. *Exceptional Children, 79*(1), 91–106. https://doi.org/10.1177/001440291207900105

Kettler, R. J., Elliott, S. N., & Beddow, P. A. (2009). Modifying achievement test items. A theory-guided and data-based approach for better measurement of what students with disabilities know. *Peabody Journal of Education, 84*(4), 529–551. https://doi.org/10.1080/01619560903240996

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Kopriva, R. (2000). *Ensuring accuracy in testing for English-language learners*. Council of Chief State School Officers. https://eric.ed.gov/?id=ED454703

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klimek, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, J., & Weiß, M. (2002). *PISA 2000: Dokumentation der Erhebungsinstrumente* [PISA 200: Documentation of the survey instruments]. Buch- und Offsetdruckerei H. Heinemann GmbH & Co.

Kutner, M., Greenberg, E., Jin, Y., Boyle, B., Hsu, Y., & Dunleavy, E. (2007). *Literacy in everyday life: Results from the 2003 national assessment of adult literacy (NCES 2007–480)*. National Center for Education Statistics. https://nces.ed.gov/Pubs2007/2007480_1.pdf

Landerl, K., & Moll, K. (2010). Comorbidity of learning disorders: Prevalence and familial
transmission. *Journal of Child Psychology and Psychiatry, 51*(3), 287–294. https://doi.org/10.1111/j.1469-7610.2009.02164.x

Lenhard, W., Lenhard, A., & Schneider, W. (2018). Ein Leseverständnistest für Erst- bis Siebtklässler–Version II [Reading Comprehension Test for Grade 1 to 7–Version II] [Technical manual]. Hogrefe.

Lovett, B. J. (2010). Extended time testing accommodations for students with disabilities: Answers to five fundamental questions. *Review of Educational Research, 80*(2), 611–638. https://doi.org/10.3102/0034654310364063

Lovett, M. W., Frijters, J. C., Steinbach, K. A., Sevcik, R. A., & Morris, R. D. (2021). Effective intervention for adolescents with reading disabilities: Combining reading and motivational remediation to improve outcomes. *Journal of Educational Psychology, 113*(4), 656–689. https://doi.org/10.1037/edu0000639

Martinello, M. (2008). Language and the performance of English language learners in math word problems. *Harvard Educational Review, 78*(2), 333–368. https://doi.org/10.17763/haer.78.2.70783570r1111t32

Mayer, R. E. (2005). Cognitive theory of multimedia learning. In Mayer, R. E. (Ed.), *The Cambridge handbook of multimedia learning* (pp. 31–48). Cambridge University Press.

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user’s guide* (8th ed.). Muthén & Muthén.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*(2), 77–101. https://doi.org/10.1037/0003-066X.51.2.77

Noble, R., Sireci, S. G., Wells, C. S., Kachchaf, R. R., Rosebery, A. S., & Wang, Y. C. (2020). Targeted linguistic simplification of science test items for English learners. *American Educational Research Journal, 57*(5), 2175–2209. https://doi.org/10.3102%2F0002831220905562

Norman, K., Caseau, D., & Stefanich, G. P. (1998). Teaching students with disabilities in inclusive science classrooms: Survey results. *Science Education, 82*(2), 127–146. https://doi.org/10.1002/sce.10066

Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is critical to scientific literacy. *Science Education, 87*(2), 224–240. https://doi.org/10.1023/s:10066

Olt, M. R. (2002). Ethics and distance education: Strategies for minimizing academic dishonesty in online assessment. *Online Journal of Distance Learning Administration, 5*(3). https://www.learntechlib.org/p/94889

Olvera Astivia, O. L., Gadermann, A., & Guhn, M. (2019). The relationship between statistical power and predictor distribution in multilevel logistic regression: A simulation-based approach. *BMC Medical Research Methodology, 19*(97), 1–20. https://doi.org/10.1186/s12874-019-0742-8

O’Reilly, T., & McNamara, D. S. (2007). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional “high-stakes” measures of high school students’ science achievement. *American Educational Research Journal, 44*(1), 161–196. https://doi.org/10.3102/0002831206298171

Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction, 19*(3), 228–242. https://doi.org/10.1016/j.learninstruc.2008.04.003

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*(4), 429–434. https://doi.org/10.1037/0022-0663.84.4.429

Paas, F., van Gog, T., & Sweller, J. (2010). Cognitive load theory: New conceptualizations, specifications, and integrated research perspectives. *Educational Psychology Review, 22*(2), 115–121. https://doi.org/10.1007/s10648-010-9133-8

Pennington, B. F., McGrath, L. M., & Peterson, R. L. (2019). *Diagnosing learning disorders* (3rd ed.). Guilford Press.

Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice, 30*(3), 10–28. https://doi.org/10.1111/j.1745-3992.2011.00207.x

Prophet, R. B., & Badebe, N. B. (2009). Language and student performance in junior secondary science examinations: The case of second language learners in Botswana. *International Journal of Science and Mathematics Education, 7*(2), 235–251. https://doi.org/10.1007/s10763-006-9058-3
R Core Team. (2009–2019). R: A language and environment for statistical computing, R Foundation for Statistical Computing. https://www.r-project.org

Rhodes, K. T., Branan-Martin, L., Morris, R. D., Romski, M., & Sevcik, R. A. (2015). Testing math or testing language? The construct validity of the KeyMath-Revised for children with intellectual disability and language difficulties. American Journal on Intellectual and Developmental Disabilities, 120(6), 542–568. https://doi.org/10.1352/1944-7558-120.6.542

Rivera, C., & Stansfield, C. W. (2004). The effect of linguistic simplification of science test items on score comparability. Educational Assessment, 9(3), 79–105. https://doi.org/10.1207/s15326977ea0903&4_1

Rose, J. (2009). Identifying and teaching children and young people with dyslexia and literacy difficulties. Department for Schools, Children, and Families. http://www.thedyslexia-spldtrust.org.uk/media/downloads/inline/the-rose-report.1294933674.pdf

Ruiz-Primo, M.-A., & Li, M. (2016). PISA Science contextualized items: The link between the cognitive demands and context characteristics of the items. RELIEVE, 22(1), 1–20. https://doi.org/10.7203/relieve.22.1.8280

Scruggs, T. E., & Lifson, S. (1984, October). Are learning disabled students “test-wise”? An inquiry into reading comprehension test items [Paper presentation]. Evaluation Network and the Evaluation Research Society, San Francisco, CA, United States. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.867.4019&rep=rep1&type=pdf

Seifert, K., & Espin, C. (2012). Improving reading of science text for secondary students with learning disabilities. Learning Disabilities Quarterly, 35(4), 236–247. https://doi.org/10.1177/0731948712444275

Siegel, M. A. (2007). Striving for equitable classroom assessments for linguistic minorities: Strategies for and effects of revising life science items. Journal of Research in Science Teaching, 44(6), 864–881. https://doi.org/10.1002/tea.20176

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Therrien, W. J., Taylor, J. C., Hosp, J. L., Kaldenberg, E. R., & Gosh, J. (2011). Science instruction for students with learning disabilities: A meta-analysis. Learning Disabilities Research & Practice, 26(4), 188–203. https://doi.org/10.1111/j.1540-5826.2011.10.1111/j.1540-5826.2011.010377.x

Thorndike, R. L., & Hagen, E. P. (1971). Cognitive Abilities Test [Technical manual]. Houghton Mifflin.

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. Journal of Educational Psychology, 98(4), 788–806. https://doi.org/10.1037/0022-0663.98.4.788

Trefil, J. (2007). Why science? Teachers College Press.

Tunmer, W., & Greaney, K. (2010). Defining dyslexia. Journal of Learning Disabilities, 43(2), 229–243. https://doi.org/10.1177/0022219409345009

Turkan, S., & Liu, O. L. (2012). Differential performance by English language learners on an inquiry-based science assessment. International Journal of Science Education, 34(15), 2343–2369. https://doi.org/10.1080/09500693.2012.705046

Wolf, M. K., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. Educational Assessment, 14(3–4), 139–159. https://doi.org/10.1080/10627190903425883

Yap, M. J., Tse, C.-S., & Balota, D. A. (2009). Individual differences in the joint effects of semantic priming and word frequency. Journal of Memory and Language, 61(3), 303–325. https://doi.org/10.1016/j.jml.2009.07.001

Yore, L. D., Bisanz, G. L., & Hand, B. M. (2003). Examining the literacy component of science literacy: 25 years of language arts and science research. International Journal of Science Education, 25(6), 689–725. https://doi.org/10.1080/095006903050518

Authors’ Note
We would like to thank Judith Keinath for her editorial support.

ORCID iD
Jan Retelsdorf [i] https://orcid.org/0000-0002-9892-9613

Supplemental Material
Supplemental material for this article is available online: https://doi.org/10.1177/00144029221094049.
Open Science Badge

For publishing their research plan prior to conducting the study and publishing data and materials, Cruz Neri and Retelsdorf received badges for pre-registration, open data, and open materials. The public content may be retrieved from https://osf.io/kta6c/.

Manuscript received July 2020; accepted March 2022.