Language in science performance: do good readers perform better?

Nadine Cruz Neri¹ · Karin Guill² · Jan Retelsdorf¹

Received: 7 March 2019 / Revised: 15 November 2019 / Accepted: 20 November 2019 /

Published online: 10 December 2019

© The Author(s) 2019

Abstract

Science performance is highly affected by students' reading comprehension. Recently, there has been a growing attention to the role of linguistic features for science performance, but findings are ambivalent, especially when looking into item word count. The aim of this study was to investigate the interaction of students' reading comprehension and item word count of given science measures on performance, controlling for students' cognitive abilities, gender, family language, and school track. The sample consisted of $N = 2051$ German students in grades 10 and 11. Students completed (scientific) literacy measures. We then applied a multilevel logistic regression to investigate the hypothesized interaction effect of reading comprehension and word count on students' science performance. The results showed a significant interaction of students' reading comprehension and word count on science performance, controlling for several covariates. Particularly students with high reading comprehension benefit from science items with increasing word count. Our findings empirically support previous research, showing that reading comprehension is crucial for science performance and enhances the interaction between reading comprehension and linguistic features of written text in science subjects. Finally, theoretical and practical implications and limitations of this study are discussed.

Keywords Reading comprehension · Science performance · Word count

✉ Nadine Cruz Neri
nadine.cruz.neri@uni-hamburg.de

Karin Guill
guill@ipn.uni-kiel.de

Jan Retelsdorf
jan.retelsdorf@uni-hamburg.de

¹ University of Hamburg, Von-Melle-Park 8, 20146 Hamburg, Germany

² Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany

In recent years, several studies highlighted the importance of reading comprehension for scientific literacy (e.g., Cromley 2009; Cromley et al. 2010; Snow 2010; Yore et al. 2004). Learning science is not only vital for people who want to pursue a scientific career, but also crucial for being an informed citizen since basic science knowledge is necessary to understand everyday life problems, such as personal health or global warming issues (National Research Council 2007; Trefil 2007).

Studies reveal that students' reading comprehension highly correlates with their science performance (Cromley 2009; O'Reilly and McNamara 2007), indicating that reading comprehension is a crucial prerequisite for science performance. Therefore, good readers usually perform better in science, whereas struggling readers often face obstacles in science performance due to their difficulties in reading comprehension, even though their scientific knowledge might equal those of good readers. Additionally, there has been some evidence that the relation of reading comprehension and science performance is significantly affected by specific text features, such as syntax or word count, revealing variation in item difficulty, when linguistic features are modified (e.g., Prophet and Badede 2009; Rivera and Stansfield 2004). However, especially the effects of item word count on science performance remain ambivalent, with studies reporting contradictory results (e.g., Bird and Welford 1995; Cassels and Johnstone 1984).

In this study, we aimed to investigate the relation between students' reading comprehension and their science performance depending on the word count per item. We assumed that an interaction between students' reading comprehension and item word count on science performance could be found: Especially students with high reading comprehension may benefit from increased item word count due to their competence to extract the substance of written text (Gibson and Levin 1975). To make sure that differences in science performance are due to reading comprehension and not just an effect of general cognitive abilities, we controlled for students' cognitive skills. Furthermore, we included other covariates into our model, that may affect science performance: gender (e.g., Ivanov and Nikolova 2010), family language (e.g., Turkan and Liu 2012), and school track (e.g., Wendt et al. 2010).

The importance of reading comprehension for scientific literacy and performance

The Organisation for Economic Co-operation and Development (OECD) refers to reading comprehension as a part of reading literacy, which is defined as "an individual's capacity to understand, use, reflect on and engage with written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society" (OECD 2013, p. 16). Reading comprehension is considered a cognitively demanding skill associated with the reading process in general, which occurs before, during, and after a written piece is read (Gernsbacher et al. 1990; Hannon and Daneman 2001; Salem 2016; Walker 1987).

Several studies show the impact of individuals' reading comprehension on academic performance in general (Cooper et al. 2014; McGee et al. 2002; Savolainen et al. 2008). Since understanding a written task is essential for solving it (Leiss et al. 2017), students need appropriate reading comprehension to process tasks, starting with the decoding of letters and words through to connecting sentences to a coherent mental representation (Francis et al. 2006; Hall et al. 2014; Perfetti 2007).

Researchers have found that reading comprehension is an important part of learning science and scientific literacy (Martin 1993; Yore et al. 2004), because science can only be constituted,

changed, and communicated with the help of language (Yore et al. 2003). Therefore, it is crucial for scientific achievement to be literate (Dempster and Reddy 2007; Härtig et al. 2015; Maerten-Rivera et al. 2010; Otero et al. 2002; Ozuru et al. 2009; Voss and Silfies 1996). The importance of reading comprehension for science measures is demonstrated by remarkable correlations of students' performance on science tasks and their reading competence. The correlations vary between 0.58 and 0.90 across studies (Cromley 2009; O'Reilly and McNamara 2007), which can be considered moderate to high correlations between two constructs (Cohen 1992). Therefore, students with low reading comprehension are in unfavorable circumstances in comparison to students with average or high reading comprehension, regarding their science performance. Students with a good science competence may be hindered by their (deficits in) reading comprehension to show their full potential in scientific measures.

Moreover, science performance does not only depend on students' reading comprehension but also on characteristics of science texts and test items (e.g., Bird and Welford 1995; Prophet and Badede 2009). Although the importance of linguistic features and their complexity for solving written tasks is undeniable, only a few studies have yet focused on the influence of specific linguistic features on science performance. There have been two major approaches pursued by researchers to investigate the effect of linguistic features on item difficulty and students' performance: the modification of specific linguistic features in science measures and secondary analyses.

Regarding the modification of linguistic features, it was pursued in several studies to simplify science items based on guidelines and recent findings (e.g., Cassels and Johnstone 1984). For example, it has been shown repeatedly that a high density of technical terms in science textbooks and test items tends to generate difficulties for students, which is linked to lower performance (e.g., Butler et al. 2004; Prenzel et al. 2002; Schmiemann 2011; Snow 2010). Other linguistic features that have been frequently shown to affect performance in various domains, such as mathematics and reading, have been the passive voice phrase (Berndt et al. 2004; Kharkwal and Stromswold 2014), the negation of sentences (e.g., Tamir 1993), pronouns (e.g., Oakhill and Yuill 1986), and many more. Several studies find significant effects when simplifying the language used in science items (e.g., Bird and Welford 1995; Kettler et al. 2012; Prophet and Badede 2009; Siegel 2007) and two meta-analyses confirm the effectiveness of linguistic simplification for English language learners (Kieffer et al. 2012; Pennock-Roman and Rivera 2011). Linguistic simplification is even able to reduce performance gaps between English language learners and native speakers (Rivera and Stansfield 2004). However, some studies also report non-significant results and only observe improvements of performance in descriptive analyses (e.g., Höttecke et al. 2018, while other studies even report contradictory findings with simplification not always working out as initially hoped (e.g., Leiss et al. 2017; Rivera and Stansfield 2004). Often, it remains unclear why simplifying the language in science test items sometimes fail to help students in solving the items.

Regarding secondary analyses, research found evidence for even more linguistic features that may generate difficulties for students resulting in lower science performance. Heppt et al. (2015) identified linguistic features such as general academic words, words with three or more syllables, and compounds as difficult for students. Interestingly, word count, complex sentence structures, and noun phrases were reported as non-significant. Dempster and Reddy (2007) found that complex sentences and unfamiliar words predicted response patterns the best: Both complex sentences and unfamiliar words seem to generate difficulties for students, resulting in lower performance. A further analysis of the items where students performed well on showed that these items tended to be shorter, while more difficult items had more prepositional phrases,

noun phrases, passive constructions, and words with multiple meanings. It remains unclear where the difficulty for students exactly arises. Stiller et al. (2016) further identified word count in the item stem and the responses as significant for item difficulty, while Prenzel et al. (2002) did not find a significant effect for word count.

While the majority of studies show consistent results for linguistic features such as technical terms (e.g., Butler et al. 2004; Prenzel et al. 2002; Snow 2010; Stiller et al. 2016), findings get a lot more ambivalent, when looking into item word count. As stated above, secondary analyses are not consistent: Some studies report significant effects for word count (Stiller et al. 2016), while others do not (Heppert et al. 2015; Prenzel et al. 2002). With regard to experimental studies, Cassels and Johnstone (1984) have shown that an increased word count with embedded clauses resulted in an overall increase of item difficulty and consequently in a decrease of students' performance on a chemistry test. In this case, it remains unclear if the effect arises from the word count or the embedded clauses, which increase sentence complexity. Prophet and Badede (2009), on the other hand, have shown that an extended reduction of overall word count does not necessarily result in higher science performance, demonstrating that an excessive reduction of word count leads to a loss of information, which in turn influences students' performance negatively. The researchers agree that for an effective simplification of science items by reducing word count, it is necessary to remove irrelevant information only. Bird and Welford (1995) also conducted a study in which they modified, or rather simplified, questions of a science exam by *inter alia* reducing the word count and substituting words, finding no statistically significant effects on students' overall performance. The effects of simplification leading to better performance only became significant when investigating non-native students. Still, it remains unclear again, if the better performance can be explained by the word count or the substitution of words. Several experimental studies in which linguistic features in science items, such as word count and sentence structure were modified, report ambivalent effects on students' performance (Höttecke et al. 2018; Rivera and Stansfield 2004; Abedi et al. 2003). However, it is not stated which factor specifically contributed the most to linguistic simplification and better students' performance.

In sum, prior investigations were able to identify multiple linguistic features that may have a significant effect on item difficulty for students. While the majority of these findings are consistent, the scientific evidence for the effect of word count remains vague. It seems apparent that word count may make a difference for students' science performance and item difficulty based on prior research and reported findings (e.g., Bird and Welford 1995; Cassels and Johnstone 1984; Stiller et al. 2016). However, it remains unclear if word count is the driving force in those effects or rather other linguistic features such as sentence structure, which is often modified simultaneously with word count.

The present investigation

In this study, we want to focus on the relation of science items' word count and students' reading comprehension due to the ambivalent findings of word count on item difficulty. We looked further into the effect of word count on item difficulty since findings have been very ambivalent (e.g., Bird and Welford 1995; Cassels and Johnstone 1984; Prophet and Badede 2009; Stiller et al. 2016). It may be possible to shed some new light on the ambivalence of findings by taking interaction effects of word count and students' reading comprehension into account. Students with good reading comprehension are proficient in extracting the main

substance of a written text (Gibson and Levin 1975). Hence, it seems possible that these proficient students are able to better extract the relevant information of a science item with a higher word count than students with less proficient reading comprehension are.

To investigate our research question, we examined three hypotheses. First, we aimed to support the finding that students' reading comprehension has a positive effect on science performance, meaning that science performance increases with higher reading comprehension (e.g., Bayat et al. 2014; Cromley 2009; O'Reilly and McNamara 2007).

Second, to examine the effect of word count, we analyzed science items with respect to their word count, hypothesizing that the word count of science items affects the probability of a correct response for respective items. To avoid confounding effects, we only included items with closed-response formats. Open-response formats were excluded due to the additionally required productive language skills (Brown and Hudson 1998) and due to the effects of response formats on performance (DeMars 2000; Härtig et al. 2015; Reardon et al. 2018). Furthermore, items with pictures were excluded, because it is known that processing texts with pictures may have favorable effects on reading comprehension (Carney and Levin 2002; Mayer 1989, 2001; Mayer and Gallini 1990).

Third, we hypothesized an interaction effect between item word count and students' reading comprehension on science performance that may explain former ambivalent findings (e.g., Cassels and Johnstone 1984; Stiller et al. 2016). In detail, we assumed that especially students with high reading comprehension benefit from increased item word count, leading to better science performance due to their better competence of extracting the information relevant for solving the science items correctly.

Finally, we controlled for (1) cognitive abilities, (2) gender, (3) family language, and (4) school track, when testing the relation between reading comprehension and item word count. First, while cognitive abilities are related in a moderate to high degree with reading comprehension (Kendeou et al. 2015; Naglieri 2001), cognitive abilities also correlate with science performance (0.46 to 0.77; Deary et al. 2007). Therefore, we included students' cognitive abilities in our model, expecting that the interaction effect of reading comprehension and word count remains significant after controlling for students' cognitive abilities. Second, we controlled for students' gender because male students tend to slightly outperform females in science measures (e.g., Ivanov and Nikolova 2010) and (b) females usually outperform male students on reading comprehension measures (e.g., Wendt et al. 2010). Third, we controlled for students' family language, because research suggests that students with a migratory background tend to show lower performances than native speakers (e.g., Heppt et al. 2015; Turkan and Liu 2012). Finally, we controlled for students' school track, because it has been suggested that students visiting the academic track tend to outperform students visiting the non-academic track in reading comprehension and science measures (e.g., Ivanov and Nikolova 2010; Wendt et al. 2010).

Methods

Sample

The data used in this study were retrieved from the larger study "Competencies and attitudes of students" (KESS, in German: *Kompetenzen und Einstellungen von Schülerinnen und Schülern*) conducted in Hamburg, Germany (for a description of KESS see Vieluf et al.

2011). KESS is a longitudinal study that started in 2003 and comprises five occasions of data collection. Even though KESS has finished in 2012, re-analyzing the KESS data is still valid: Neither the science curriculum in Hamburg, Germany, has changed considerably since the KESS study, nor have the tests used in educational assessment.

In the current study, we used data from the fourth data collection in 2009, where $N = 13,328$ teenagers at the end of grade 10 or the beginning of grade 11 were tested. Students, who visited the comprehensive school (non-academic track), were tested in grade 10 before the summer break, while students visiting the German Gymnasium (academic track) were tested shortly after the summer break in grade 11. If students were tested twice due to a change of school over the summer break, only their first assessment was incorporated in the data and analyzed.

The analytic sample for the current study was limited to those students who worked on the science tests; students not working on these tests due to the test design were excluded. Students that did not take part in the reading comprehension and cognitive ability tests were also excluded. Consequently, there were no missing data in the sample in the assessed competence measures due to the specific selection of our analytic sample. Missing answers on single items were coded as false. The final sample consisted of 2051 students (age: $M = 16.5$, $SD = 0.53$; 51.9% female) from 52 schools in Hamburg, Germany, which were tested in the non-academic track in grade 10 (43.3%) or in the academic track in grade 11 (56.7%). Most students only spoke German at home (84.7%), while a minority used German and another language or just another language at home (15.3%). Students' demographics were obtained by questionnaires. All measures were assessed in German.

Measures

Science The science test was based on the scientific-literacy concept, which measures the competence to apply scientific knowledge appropriately and accordingly to a specific situation (cf. Vieluf et al. 2011). The tests for grades 10 and 11 consisted of 46 items in total. The items stem from the Third International Mathematics and Science Study conducted in 1995 (TIMSS/III; Baumert et al. 2000).

For our study and its specific research question, items with pictures connected to the individual task ($n = 11$) and with an open-answer format ($n = 7$) were excluded from the analyses due to possible confounding of additional factors, in particular required productive skills and additional effects on reading comprehension (e.g., Brown and Hudson 1998; Mayer 1989, 2001). Pictures in science items have been found to affect science performance, depending on the type of picture. Decorative pictures, for example, do not affect science performance, but other types of pictures, such as transformational ones, can enhance systematic thinking, which in turn, may affect performance positively (e.g., Levin et al. 1987). The remaining 28 items were multiple-choice items with three to five alternative answers, including items from different subjects: biology ($n = 9$), chemistry ($n = 5$), earth sciences ($n = 6$), environmental issues and the nature of science ($n = 3$), and physics ($n = 5$).

The science items were coded by two independent raters with regard to their word count. For the coding, the number of words in the item, including the question and the multiple-choice answers, was counted and summarized ($M = 36.9$, $SD = 20.8$). Cohen's kappa was 0.95 indicating an excellent degree of agreement of the coded word count (Cohen 1960). The slight deviation in coding was a result of initial uncertainties regarding symbols in the said items. Moreover, the science items in KESS were relatively simple in sentence structure: There were almost no embedded or subordinate clauses with the sentence structure mainly consisting of

main clauses. Negations, pronouns, noun phrases, and nominalizations were kept to a minimum.

All students in our analytic sample completed the 28 items. Students' answers on each item were coded as correct or false: Missing answers on single items were coded as false. In a Rasch model for dichotomous data, weighted likelihood estimates (WLEs; Warm 1989) were estimated as students' science performance scores and standardized ($M=100$, $SD=30$), using ConQuest (Wu et al. 1998). In ConQuest, item and person parameters are estimated with the marginal maximum likelihood (MML) method, which provides reliable estimates (Rost 2004). The WLE reliability of the science performance was sufficient (0.72).

The WLEs were primarily used to calculate the correlations of students' science performance and reading comprehension and cognitive abilities, respectively. To investigate the research question, students' answers on each item (0 = false answer, 1 = correct answer) were used in the statistical analyses.

Reading comprehension The assessment of students' reading comprehension consisted of eight texts (two factual texts, two stories, three newspaper articles, and one text composed of graphics) with 58 tasks, using mostly multiple-choice answer formats with four alternative answers and a few open-answer formats. The tasks were rotated systematically between the booklets with a multi-matrix design. Therefore, the booklets varied in their number of items, ranging from 36 to 39 items, all of them including 24 anchor-items.

WLEs (Warm 1989) were estimated as students' reading comprehension, using ConQuest (Wu et al. 1998). The WLE reliability of the reading comprehension tasks was good (0.82).

Cognitive abilities Students completed a non-verbal subtest of a German test to assess children's and adolescents' cognitive abilities (in German: *Kognitiver Fähigkeitstest*; Heller and Perleth 2000), which is based on Thorndike and Hagen's Cognitive Abilities Test (Thorndike and Hagen 1971). According to Heller and Perleth (2000), the construct validity is given with the non-verbal subtest and as per test manual, the internal consistency is 0.93, indicating an excellent estimate of reliability of the test scores. Students had 8 min to process the subtest, which consists of 25 items. This subtest is considered to be a fair indicator of general cognitive abilities (Neisser et al. 1996).

Since cognitive abilities have been tested at neither the end of grade 10 nor the beginning of grade 11, we used students' performance from the previous data collection in grade 8. Using ConQuest (Wu et al. 1998), WLEs (Warm 1989) were estimated for students' cognitive abilities. The WLE reliability of the cognitive ability test was good (0.82).

Statistical analyses

We analyzed how the interaction of word count and students' reading comprehension affects the probability of a correct response on science items applying multilevel logistic regression modeling using the software *Mplus* 8.3 (Muthén and Muthén 1998-2015). We conducted a multilevel regression model due to the structure and hierarchy of the given data. We applied multilevel modeling since we considered the science items with varying word count as being nested in students. Furthermore, we applied logits regression in our multilevel model due to the binary nature of our dependent variable science performance on item level (0 = false answer, 1 = correct answer). Moreover, in our multilevel model, the item characteristic word count was

treated as a within-level variable (on level 1). Person characteristics, i.e., students' reading comprehension and other covariates (cognitive abilities, gender, family language, and school track), were treated as between-level variables (on level 2).

Finally, to test the effect of the interaction between students' reading comprehension and items' word count on science performance, a cross-level interaction between these variables was included in our model. To test for the role of cognitive abilities, we further included a cross-level interaction between students' cognitive abilities and item word count on science performance. Furthermore, we controlled for three dummy-coded variables: gender (0 = male, 1 = female), family language (0 = German, 1 = family language German and another language or only other than German), and school track (which equals the time of assessment; 0 = non-academic track, 1 = academic track). All predictors but the dummy-coded variables on within- and between-level were standardized ($M = 0$, $SD = 1$) for an easier interpretation of the effects and odds ratios were calculated for the effects.

Results

Descriptive statistics and correlations

Since measuring science performance involves knowledge from five domains (biology, chemistry, earth sciences, environmental issues and the nature of science, and physics), we tested if the according items differed in their word count. We did not find significant differences, $F(4,23) = 0.44$, $p = 0.777$. The probability of a correct response for the items measuring science performance ranged from 0.21 to 0.93 ($M = 0.56$, $SD = 0.16$).

The correlations between students' science performance, reading comprehension, and cognitive abilities are presented in Table 1. All three variables correlated positively with moderate to large effect sizes. Table 1 further contains the correlations of these variables with the remaining covariates gender, family language, and school track.

Moreover, we tested differences between students from the academic vs. non-academic track. Academic track students outperformed non-academic track students in science performance ($d = 1.08$), reading comprehension ($d = 1.33$), and cognitive abilities ($d = 0.88$). Finally, the ICC for the science items indicated that 10.3% of the variance in science performance is due to differences between students.

Multilevel analyses

The results of our multilevel model, including the unstandardized coefficient (B), the standard error (SE), the p value, the odds ratio (OR), and the confidence intervals (CI), are presented in Table 2. Contrary to our prediction, the item word count was not a significant predictor on the within-level (level 1). However, reading comprehension and cognitive abilities proved to be significant predictors on the between-level (level 2), indicating a higher probability of a correct response with increased reading comprehension and cognitive abilities. Moreover, the remaining covariates (gender, family language, and school track) also were significant predictors on the between-level.

These main effects, however, need to be interpreted in light of the interaction between reading comprehension and word count, which was significant. Figure 1 portrays the interaction effect (black line) and the 95% confidence intervals (gray line). The x -axis of the plot

Table 1 Correlations among students' characteristics (between-level variables)

Variable	1	2	3	4	5
1. Reading comprehension	—				
2. Cognitive abilities	0.40**	—			
3. Science performance	0.59**	0.43**	—		
4. Gender	0.11**	0.02**	-0.10**	—	
5. Family language	0.25**	0.16**	0.26**	0.01*	—
6. School track	0.55**	0.40**	0.47**	0.06**	0.24**

Students' ability scores were weighted likelihood estimates (WLEs). The WLE for science was not used in our regression models, but only to present a more general impression of the correlation of reading comprehension, cognitive abilities, and science. The following variables were dummy-coded: gender (0 = male, 1 = female), family language (0 = German, 1 = family language German, and another language or only other than German), and school track (0 = non-academic track, 1 = academic track). * $p < 0.05$; ** $p < 0.01$

depicts students' range of values for their reading comprehension from -2 SD to $+2$ SD, while the y -axis depicts the range of values for the conditional slope of science performance on items' word count ($B_{\text{science performance on word count}}$). The x -axis depicts standardized values ($M = 0$, $SD = 1$) while the y -axis depicts the unstandardized regression coefficient B . The increasing black line of the interaction effects shows that students with higher reading comprehension benefitted from increased item word count, which led to increased science performance. However, we also need to take the confidence intervals into account, because they help understand on which level of reading comprehension the interaction effect on science performance is significant (also known as region of significance, cf. Johnson and Neyman 1936). In Fig. 1, you can see that the confidence intervals do not cross the x -axis anymore just over -1 SD, which indicates that the interaction effect becomes significant at that point. This means that particularly for students with very low reading comprehension levels (below -1 SD), there are no effects of word count on science performance. For students with higher reading comprehension, the interaction effect became significant, which means that they scored significantly higher on science items with higher word count.

Table 2 Results of the multilevel analyses predicting science performance (unstandardized parameters)

Predictor	B	SE	p	OR	95% CI
Intercept	-0.113	0.036	0.002		[-0.183, -0.043]
Level 1					
Word count	0.038	0.026	0.141	1.038	[-0.028, 0.088]
Level 2					
Reading comprehension	0.314	0.021	< 0.001	1.369	[0.274, 0.355]
Cognitive abilities	0.136	0.015	< 0.001	1.146	[0.107, 0.165]
Gender	-0.238	0.024	< 0.001	0.788	[-0.286, -0.191]
Family language	0.188	0.034	< 0.001	1.207	[0.121, 0.255]
School track	0.227	0.030	< 0.001	1.255	[0.167, 0.286]
Interaction effects					
Reading comprehension word count	0.047	0.0121	< 0.001	1.048	[0.024, 0.070]
Cognitive abilities word count	0.029	0.010	0.004	1.029	[0.009, 0.048]
Gender $\times$ word count	0.041	0.017	0.018	1.042	[0.007, 0.074]
Family language $\times$ word count	0.004	0.024	0.883	1.004	[-0.44, 0.051]
School track $\times$ word count	0.022	0.021	0.297	1.022	[-0.019, 0.063]

The following variables were dummy-coded: gender (0 = male, 1 = female), family language (0 = German, 1 = family language German, and another language or only other than German), and school track (0 = non-academic track, 1 = academic track)

Fig. 1 Interaction of reading comprehension and word count on science performance (black line). The y-axis depicts a range of values for the conditional slope of science performance on word count, while the x-axis depicts students' range of values for reading comprehension from -2 SD to $+2$ SD. The plot also includes the 95% confidence interval (gray line)

Discussion

The aim of this research was to investigate the effect of the relation between the word count in science items and students' reading comprehension on science performance. We expected the interaction of reading comprehension and word count of science items to affect science performance and examined our hypothesis by applying a multilevel logistic regression model drawing on a relatively large student sample. Our hypothesis was corroborated: The interaction between word count and reading comprehension affected science performance significantly even when controlling for students' cognitive abilities, gender, family language, and school track. In detail, the interaction indicated that particularly students with good reading comprehension benefitted from increased item word count in the investigated science measures, while there are no interaction effects on students with low reading comprehension.

One could argue that the interaction is relatively small. However, it needs to be taken into consideration that interaction effects are difficult to detect in non-experimental studies, because they are usually quite small in the psychological research field (Champoux and Peters 1987; Chaplin 1991) and due to other reasons concerning assessment, such as unreliability of measures and the lack of pronounced profiles (e.g., Busemeyer and Jones 1983; McClelland and Judd 1993). Large sample sizes, such as in our analytic sample, help to detect these interaction effects if they truly exist (Trautwein et al. 2012) and, therefore, even small interaction effects can be important and should be interpreted in the research context (Chaplin 1991).

Theoretical and practical implications

Our findings contribute to the growing body of literature showing the importance of reading comprehension for science performance (e.g., Hall et al. 2014; O'Reilly and McNamara 2007). Furthermore, the significant interaction between reading comprehension and word count underlines that linguistic item characteristics may not have the same consequences for different test takers, but vary as a function of the test takers' prerequisites. Concerning the assessment of science performance, this result may be interpreted from at least two perspectives. First, one may argue that tests

measuring science performance should not depend too strongly on the test takers' language skills, since this might systematically disadvantage particular social groups, such as non-native speakers (Bird and Welford 1995; Lee 2005; Prophet and Badede 2009) and thus may be seen as a serious threat to test validity. Second, language skills are often discussed as something inherent to scientific literacy (e.g., Härtig et al. 2015; Yore et al. 2004) and, thus, language skills may also be seen as an integral part of science performance measures.

Regardless of the implications for science performance assessment, our findings demonstrate once again the importance of reading comprehension and linguistic features in students' academic life. While good readers benefit from their reading skills in other subjects, weak readers face obstacles not only in their linguistic competencies but also in other domains like science. Hence, to improve the science performance especially of weak readers, reading comprehension and reading strategies need to be trained, so weak readers do not continue to suffer from their below-average reading comprehension that may lay the foundations for their general academic achievement (e.g., Cooper et al. 2014; McGee et al. 2002; Savolainen et al. 2008). Consequently, reading comprehension and strategies to deal with difficulty-generating linguistic features should not only be taught in regular language classes, but should also be promoted in science education (or indeed throughout all classes). Providing instructional scaffolding techniques, for example, has been shown to be effective in improving students' reading comprehension (Huggins and Edwards 2011; Salem 2016). These instructional scaffolds work as a temporary guidance and support for students and are removed little by little over time, allowing students to acquire new skills and independence from their teachers, gaining freedom in their way of learning (Salem 2016).

Furthermore, the finding that increased item word count shows an increase in students' science performance for good readers sheds some light on the ambivalent reports of previous research. First, unlike Bird and Welford (1995), Cassels and Johnstone (1984), and Prophet and Badede (2009), we found that particularly good readers benefitted from increased word count that led to decreased item difficulty, which is in line with findings from Stiller et al. (2016). This might be due to the fact that longer tasks provide more information for students to solve the science items, which particularly good readers are able to benefit from by recognizing and utilizing the additionally allocated information (McNamara and O'Reilly 2010; Lau and Chan 2003).

Second, another explanation could be the key to the ambivalence of item word counts' importance: Irrelevant information (Cassels and Johnstone 1984; Prophet and Badede 2009) is sometimes included in academic achievement measures to distract students and to test their competence in capturing the essence of a task, being able to recognize and neglect irrelevant information. It may be the case that only students with high reading comprehension are able to do so, whereas students with lower reading comprehension fail to recognize relevant information that is embedded in irrelevant information.

Third, contextualized items may further explain our results. Contextual items are items that contain "supplemental information that precedes or follows the item question, such as a description of a lab setup, a natural phenomenon, or a practical problem" (Ruiz-Primo and Li 2016, p. 2). It has been argued that contextual items provide students with a context that is helpful to solve the task, because it makes test items more concrete and, therefore, realistic and relevant (Haladyna 1997). Some guidelines even suggest providing contextual information within an item, because it may be helpful for students (e.g., Kopriva 2000). Most science items in KESS are embedded in a context to test if students are able to apply their scientific knowledge into real-life contexts. Naturally, word count increases if a context is provided

for an item. Considering our results, it seems possible that students with high reading comprehension benefit from items with an increased word count, because they benefit from the context provided in the item. Simultaneously, students with low reading comprehension may not be able to benefit from these contextualized items. While this reasoning seems plausible, Ruiz-Primo and Li (2016) state that much more research is needed in order to understand the underlying processes of dealing with contextual items since to date little is known about the link of contextual items and students' performance.

To the best of our knowledge, former studies did not take into consideration that item word count in science items might affect science performance significantly depending on students' reading comprehension. While these studies usually focused on either reading comprehension or item word count (and further linguistic features), our study combined these two variables.

Limitations and future directions

Despite having some notable strengths, such as the large sample of students, this study also has at least three noteworthy limitations that need to be taken into consideration when interpreting our results. First, the study was not conducted with an experimental design, and hence, causal conclusions cannot be made. Future research should focus on systematically modifying linguistic features in science measures. Having said that, the tests and the setting in this study were part of an existing large-scale assessment study, so that one benefit of our research might be its high ecological validity for real test situations. Thus, our results complement findings from experimental research (e.g., Bird and Welford 1995; Cassels and Johnstone 1984; Höttecke et al. 2018; Prophet and Badede 2009), drawing a more complete picture of language in science education.

Second, we focused on word count as one particular linguistic feature that may contribute to a variation in item difficulty to shed some light on the ambivalent findings (e.g., Bird and Welford 1995; Stiller et al. 2016). However, previous research indicated that further linguistic features such as academic language, tenses, or negation may also affect science performance (Bird and Welford 1995; Cassels and Johnstone 1984; Höttecke et al. 2018; Prophet and Badede 2009). Having said that, in this study, we were somehow restricted in the investigation of various linguistic features since the data and science tasks were given due to secondary data analyses. Coding the science tasks with regard to several linguistic features proved itself to be difficult since there was not enough linguistic variation in the available science measures.

Third, the question arises if the results can be generalized onto other student populations. KESS was conducted in Hamburg, Germany, assessing all students from a particular cohort. The test versions to assess students' competencies were rotated by chance, and hence, it seems plausible that our analytic sample is representative for other students in Hamburg, Germany. We cannot ensure generalization of our analytic sample onto other students in other federal states, but due to quite similar curricula in all federal states in Germany, it seems plausible that our results can be generalized to students in Germany.

Conclusion

Our study complements previous research, demonstrating that students' reading comprehension and item word count as one linguistic feature affect science performance. Our results indicate that with increased reading comprehension and word count, students' science performance increases as well. Particularly good readers seem to be able to extract the substance of written text from more extensive tasks and utilize it to solve items correctly. To ensure that

weak readers do not suffer from a lack of competence in one area in other domains, enhancing reading comprehension in schools should be seen as a task not only for language education but also for science education or even across all subjects. Further research is needed, testing these findings in an experimental design and considering further linguistic features.

Acknowledgments This paper uses data from the longitudinal study KESS. The data set was generated by the Free and Hanseatic City of Hamburg through the Ministry of Schools and Vocational Training between 2003 and 2012. It has been provided to the MILES scientific consortium (Methodological Issues in Longitudinal Educational Studies) for a limited period with the aim of conducting in-depth examinations of scientific questions. MILES is coordinated by the Leibniz Institute for Science and Mathematics Education (IPN). We would like to thank Judith Keinath for the editorial assistance with this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abedi, J., Courtney, M., & Leon, S. (2003). *Effectiveness and validity of accommodations for English language learners in large-scale assessments* (CSE Report No. 608). Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.
- Baumert, J., Bos, W., & Lehmann, R. (2000). *TIMSS/III Dritte Internationale Mathematik- und Naturwissenschaftsstudie – mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn* [TIMSS/III The Third International Mathematics and Science Study – Mathematical and scientific education at the end of school career]. Wiesbaden: Springer VS. <https://doi.org/10.1007/978-3-322-83411-9>.
- Bayat, N., Sekercioglu, G., & Bakir, S. (2014). The relationship between reading comprehension and success in science. *Education in Science*, 39, 457–466. <https://doi.org/10.15390/EB.2014.3693>.
- Berndt, R. S., Mitchum, C. C., Burton, M. W., & Haendiges, A. N. (2004). Comprehension of reversible sentences in aphasia: the effects of verb meaning. *Cognitive Neuropsychology*, 21, 229–244. <https://doi.org/10.1080/02643290342000456>.
- 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, 389–397. <https://doi.org/10.1080/0950069950170309>.
- Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *Teachers of English to Speakers of Other Languages*, 32, 653–675. <https://doi.org/10.2307/3587999>.
- Busemeyer, J. R., & Jones, L. E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin*, 93, 549–562. <https://doi.org/10.1037//0033-2909.93.3.549>.
- Butler, F. A., Bailey, A.L., Stevens, R., Huang, B., & Lord, C. (2004). *Academic English in fifth-grade mathematics, science and social studies textbooks*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Carney, R. N., & Levin, J. R. (2002). Pictorial illustrations still improve students' learning from text. *Educational Psychology Review*, 14, 5–26. <https://doi.org/10.1023/A:1013176309260>.
- Cassels, J. R. T., & Johnstone, A. H. (1984). The effect of language on student performance on multiple choice tests in chemistry. *Journal of Chemical Education*, 61, 613–615. <https://doi.org/10.1021/ed061p613>.
- Champoux, J. E., & Peters, W. S. (1987). Form, effect size and power in moderated regression analysis. *Journal of Occupational Psychology*, 60(3), 243–255. <https://doi.org/10.1111/j.2044-8325.1987.tb00257.x>.
- Chaplin, W. F. (1991). The next generation of moderator research in personality psychology. *Journal of Personality*, 59(2), 143–178 Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/1880698>.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1, 37–46. <https://doi.org/10.1177/001316446002000104>.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 115–159. <https://doi.org/10.1037/0033-2909.112.1.155>.

- 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, 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, 89–118. <https://doi.org/10.1080/02702710802274903>.
- Cromley, J. G., Snyder-Hogan, L. E., & Luciw-Dubas, U. A. (2010). Reading comprehension of scientific text: a domain-specific test of the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, 102, 687–700. <https://doi.org/10.1037/a0019452>.
- Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, 35, 13–21. <https://doi.org/10.1016/j.intell.2006.02.001>.
- DeMars, C. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, 13, 55–77. https://doi.org/10.1207/s15324818ame1301_3.
- Dempster, E. R., & Reddy, V. (2007). Item readability and science achievement in TIMSS 2003 in South Africa. *Science Education*, 91, 906–925. <https://doi.org/10.1002/sce.20225>.
- Francis, D. J., Snow, C. E., August, D., Carlson, C. D., Miller, J., & Iglesias, A. (2006). Measures of reading comprehension: a latent variable analysis of the diagnostic assessment of reading comprehension. *Scientific Studies of Reading*, 10, 301–322. https://doi.org/10.1207/s1532799xssr1003_6.
- Gemsbacher, M. A., Vamer, K. R., & Faust, M. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(3), 430–445. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4301443/>.
- Gibson, E. J., & Levin, H. (1975). *The psychology of reading*. Cambridge: MIT Press.
- Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Boston: Allyn and Bacon.
- Hall, S. S., Kowalski, R., Paterson, K. B., Basran, J., Filik, R., & Maltby, J. (2014). Local text cohesion, reading ability and individual science aspirations: key factors influencing comprehension in science classes. *British Educational Research Journal*, 41, 122–142. <https://doi.org/10.1002/berj.3134>.
- Hannon, B., & Daneman, M. (2001). A new tool for measuring and understanding individual differences in the component processes of reading comprehension. *Journal of Educational Psychology*, 93, 103–128. <https://doi.org/10.1037/0022-0663.93.1.103>.
- 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, 763–779. <https://doi.org/10.1007/s11618-015-0646-2>.
- Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision: KFT 4–12+ R [Cognitive Ability Test]*. Göttingen: 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, 61–82. <https://doi.org/10.1002/rq.83>.
- Höttecke, D., Feser, M. S., Heine, L., & Ehmke, T. (2018). Do linguistic features influence item difficulty in physics assessments? *Science Education Review Letters*, 2018, 1–6. <https://doi.org/10.18452/19188>.
- Huggins, G. E., & Edwards, R. (2011). Scaffolding to improve reading comprehension and to write scholarly research sciences. *Illinois Reading Council Journal*, 1(16), 30–36.
- Ivanov, S., & Nikolova, R. (2010). Naturwissenschaftliche Kompetenzen. In W. Bos & C. Gröhlisch (Eds.), *Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 8 (pp. 79–99) [Competencies and attitudes of students at the end of grade 8]*. Münster: Waxmann.
- Johnson, P. O., & Neyman, J. (1936). Tests of certain linear hypotheses and their applications to some educational problems. *Statistical Research Memoirs*, 1, 57–93.
- Kendeou, P., Papadopoulos, T. C., & Spanoudis, G. (2015). Reading comprehension and PASS Theory. In T. C. Papadopoulos, R. K. Parrila, & J. R. Kirby (Eds.), *Cognition, intelligence and achievement* (pp. 117–136). London: Academic Press. <https://doi.org/10.1016/C2012-0-07388-7>.
- 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. Retrieved from <https://eric.ed.gov/?id=EJ979576>.
- Kharkwal, G., & Stromswold, K. (2014). Good-enough language processing: evidence from sentence-video matching. *Journal of Psycholinguistic Research*, 43(1), 27–43. <https://doi.org/10.1007/s10936-013-9239-5>.
- Kieffer, M. J., Rivera, C., & Francis, D. J. (2012). *Practical guidelines for the education of English language learners. Research-based recommendations for the use of accommodations in large-scale assessments. 2012 update*. Portsmouth, NH: RMC Research Corporation, Center on Instruction.

- Kopriva, R. (2000). *Ensuring accuracy in testing for English-language learners*. Washington, DC: Council of Chief State School Officers Retrieved from <https://eric.ed.gov/?id=ED454703>.
- Lau, K. L., & Chan, D. W. (2003). Reading strategy use and motivation among Chinese good and poor readers in Hong Kong. *Journal of Research in Reading*, 26, 177–190. <https://doi.org/10.1111/1467-9817.00195>.
- Lee, O. (2005). Science education with English language learners: synthesis and research agenda. *Review of Educational Research*, 75, 491–530. <https://doi.org/10.3102/00346543075004491>.
- Leiss, D., Domenech, M., Ehmke, T., & Schwippert, K. (2017). Schwer - schwierig - diffizil: Zum Einfluss sprachlicher Komplexität von Aufgaben auf fachliche Leistungen in der Sekundarstufe I [on the influence of tasks' language complexity on performance in secondary level I]. In D. Leiss, M. Hagena, A. Neumann, & K. Schwippert (Eds.), *Mathematik und Sprache. Empirischer Forschungsstand und unterrichtliche Herausforderungen* (pp. 99–125). Münster: Waxmann.
- Levin, J. R., Anglin, G. J., & Carney, R. N. (1987). On empirically validating functions of pictures in prose. In D. M. Willows & H. A. Houghton (Eds.), *The psychology of illustration: I. basic research* (pp. 51–85). New York: Springer. https://doi.org/10.1007/978-1-4612-4674-9_2.
- Maerten-Rivera, J., Myers, N., Lee, O., & Penfield, R. (2010). Student and school predictors of high-stakes assessment in science. *Science Education*, 94, 937–962. <https://doi.org/10.1002/sce.20408>.
- Martin, J. R. (1993). Literacy in science: learning to handle text as technology. In M. A. K. Halliday & J. R. Martin (Eds.), *Writing science: Literacy and discursive power* (pp. 166–202). Washington, DC: Falmer Press.
- Mayer, R. E. (1989). Systematic thinking fostered by illustrations in scientific text. *Journal of Educational Psychology*, 81, 240–246. <https://doi.org/10.1037/0022-0663.81.2.240>.
- Mayer, R. E. (2001). *Multimedia learning*. New York: Cambridge University Press.
- Mayer, R. E., & Gallini, J. K. (1990). When is an illustration worth ten thousand words? *Journal of Educational Psychology*, 82, 715–726. <https://doi.org/10.1037/0022-0663.82.4.715>.
- McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, 114(2), 376–390. <https://doi.org/10.1037/0033-2909.114.2.376>.
- McGee, R., Prior, M., Williams, S., Smart, D., & Sanson, A. (2002). The long-term significance of teacher-rated hyperactivity and reading ability in childhood: findings from two longitudinal studies. *Journal of Child Psychology and Psychiatry*, 43(8), 1004–1017. <https://doi.org/10.1111/1469-7610.00228>.
- McNamara, D. S., & O'Reilly, T. (2010). Theories of comprehension skill: knowledge and strategies versus capacity and suppression. In A. M. Columbus (Ed.), *Advances in psychology research* (pp. 113–136). New York: Nova Science Publishers.
- Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide* (Seventh ed.). Los Angeles, CA: Muthén & Muthén.
- Naglieri, J. A. (2001). Do ability and reading achievement correlate? *Journal of Learning Disabilities*, 34, 304–305. <https://doi.org/10.1177/002221940103400403>.
- National Research Council (2007). *Rising above the gathering storm: energizing and employing. America for a brighter future*. Washington, DC: The National Academies Press. <https://doi.org/10.17226/11463>.
- 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. J., & Urbina, S. (1996). Intelligence: knowns and unknowns. *American Psychologist*, 51, 77–101. <https://doi.org/10.1037/0003-066X.51.2.77>.
- O'Reilly, T., & McNamara, D. S. (2007). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional highstakes measures of high school students science achievement. *American Educational Research Journal*, 44, 161–196. <https://doi.org/10.3102/0002831206298171>.
- Oakhill, J. V., & Yuill, N. M. (1986). Pronoun resolution in skilled and less-skilled comprehenders: effects of memory load and inferential complexity. *Language and Speech*, 29, 25–37. <https://doi.org/10.1177/002383098602900104>.
- OECD. (2013). PISA 2012 assessment and analytical framework: mathematics, reading, science, problem solving and financial literacy. *OECD Publishing*. <https://doi.org/10.1787/9789264190511-en>.
- Otero, J., León, J., & Graesser, A. (2002). *The Psychology of science text comprehension*. Mahwah, NJ: Erlbaum. <https://doi.org/10.4324/9781410612434>.
- 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, 228–242. <https://doi.org/10.1016/j.learninstruc.2008.04.003>.
- 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, 10–28. <https://doi.org/10.1111/j.1745-3992.2011.00207.x>.
- Perfetti, C. A. (2007). Reading ability: lexical quality to comprehension. *Scientific Studies of Reading*, 11, 357–383. <https://doi.org/10.1080/1088430701530730>.
- Prenzel, M., Häußler, P., Rost, J., & Senkbeil, M. (2002). The PISA science test: can we predict the item difficulties? *Unterrichtswissenschaft*, 30, 120–135 Retrieved from https://www.pedocs.de/volltexte/2013/7682/pdf/UnterWiss_2002_1_Prenzel_et_al_PISA_Naturwissenschaftstest.pdf.

- Prophet, R. B., & Badede, 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, 235–251. <https://doi.org/10.1007/s10763-006-9058-3>.
- Reardon, S. F., Kalogrides, D., Podolsky, A., & Zárate, R. C. (2018). The relationship between test item format and gender achievement gaps on math and ELA tests in fourth and eighth grades. *Educational Reseracher*, 47, 284–294. <https://doi.org/10.3102/0013189X18762105>.
- Rivera, C., & Stansfield, C. W. (2004). The effect of linguistic simplification of science test items on score comparability. *Educational Assessment*, 9, 79–105. https://doi.org/10.1207/s15326977ea0903&4_1.
- Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* [textbook test theory – Test construction]. Bern: Hans Huber.
- 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, 1–20. <https://doi.org/10.7203/relieve.22.1.8280>.
- Salem, A. A. M. S. (2016). Scaffolding reading comprehension skills. *English Language Teaching*, 10, 97–111. <https://doi.org/10.5539/elt.v10n1p97>.
- Savolainen, H., Ahonen, T., Aro, M., Tolvanen, A., & Holopainen, L. (2008). Reading comprehension, word reading and spelling as predictors of school achievement and choice of secondary education. *Learning and Instruction*, 18, 201–210. <https://doi.org/10.1016/j.learninstruc.2007.09.017>.
- Schmiemann, P. (2011). Terminology in biological test items. *Zeitschrift für Didaktik der Naturwissenschaften*, 17, 115–136 Retrieved from http://archiv.ipn.uni-kiel.de/zfdn/pdf/17_Schmiemann.pdf.
- Siegel, M. A. (2007). Striving for equitable classroom assessments for linguistic minorities. *Journal of Research in Science Teaching*, 44, 864–881. <https://doi.org/10.1002/tea.20176>.
- Snow, C. E. (2010). Academic language and the challenge of reading for learning about science. *Science*, 328, 448–450. <https://doi.org/10.1126/science.1182596>.
- Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V., Krüger, D., & Upmeyer Zu Belzen, A. (2016). Assessing scientific reasoning: a comprehensive evaluation of item features that affect item difficulty. *Assessment & Evaluation in Higher Education*, 41, 721–732. <https://doi.org/10.1080/02602938.2016.1164830>.
- Tamir, P. (1993). Positive and negative multiple choice items: how different are they? *Studies in Educational Evaluation*, 19(3), 311–325 Retrieved from <https://eric.ed.gov/?id=EJ471898>.
- Thorndike, R. L., & Hagen, E. P. (1971). *Cognitive abilities test [Measurement instrument]*. Boston: Houghton Mifflin.
- Trautwein, U., Marsh, H. W., Nagengast, B., Lüdtke, O., Nagy, G., & Jonkmann, K. (2012). Probing for the multiplicative term in modern expectancy–value theory: a latent interaction modeling study. *Journal of Educational Psychology*, 104, 763–777. <https://doi.org/10.1037/a0027470>.
- Trefil, J. (2007). *Why science?* New York: Teachers College Press.
- Turkan, S., & Liu, O. L. (2012). Differential performance by English language learners on an inquiry-based science assessment. *International Journal of Science Education*, 34, 2343–2369. <https://doi.org/10.1080/09500693.2012.705046>.
- Vieluf, U., Ivanov, S., & Nikolova, R. (2011). *Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen am Ende der Sekundarstufe I und zu Beginn der gymnasialen Oberstufe* [Competencies and attitudes of students at the end of secondary level I and at the beginning of high school]. Münster: Waxmann.
- Voss, J., & Silfies, L. (1996). Learning from history text: the interaction of knowledge and comprehension skill with text structure. *Cognition and Instruction*, 14, 45–68. https://doi.org/10.1207/s1532690xc1401_2.
- Walker, C. H. (1987). Relative importance of domain knowledge and overall aptitude on acquisition of domain related information. *Cognition and Instruction*, 4, 25–42. https://doi.org/10.1207/s1532690xc140401_2.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427–450. <https://doi.org/10.1007/BF02294627>.
- Wendt, H., Gröhlich, C., Guill, K., Scharenberg, K., & Bos, W. (2010). Die Kompetenzen der Schülerinnen und Schüler im Leseverständnis. In W. Bos & C. Gröhlich (Eds.), *KESS 8. Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 8* (pp. 21–36) [Competencies and attitudes of students at the end of grade 8]. Münster: Waxmann.
- Wu, M. L., Adams, R. J., & Wilson, M. (1998). ConQuest: Generalized item response modeling software [computer software]. Melbourne: Australian Council for Educational Research.
- 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, 689–725. <https://doi.org/10.1080/09500690305018>.
- Yore, L. D., Hand, B., Goldman, S. R., Hildebrand, G. M., Osborne, J. F., Treagust, D. F., & Wallace, C. S. (2004). New directions in language and science education research. *Reading Research Quarterly*, 39, 347–352. <https://doi.org/10.1598/RRQ.39.3.8>.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Nadine Cruz Neri. University of Hamburg, Von-Melle-Park 8, 20146 Hamburg, Germany. E-mail: nadine.cruz.neri@uni-hamburg.de

Current themes of research:

Science text comprehension. Math text comprehension.

Most relevant publications in the field of Psychology of Education:

No previous publications

Karin Guill. Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany. E-mail: guill@ipn.uni-kiel.de

Current themes of research:

Private tutoring. Parental support. Academic tracking. Transition points in educational systems.

Most relevant publications in the field of Psychology of Education:

Guill, K. & Lintorf, K. (2018). Private tutoring when stakes are high: insights from the transition from primary to secondary school in German. *International Journal of Educational Development*. Online first. doi: 10.1016/j.ijedudev.2018.08.001

Guill, K., Lüdtke, O. & Köller, O. (2017). Academic tracking is related to gains in students' intelligence over four years: evidence from a propensity score matching study *Learning and Instruction*, 47, 43–52. doi: <https://doi.org/10.1016/j.learninstruc.2016.10.001>.

Guill, K. & Bos, W. (2014). Effectiveness of private tutoring in mathematics with regard to subjective and objective indicators of academic achievement. Evidence from a German secondary school sample. *Journal for Educational Research Online*, 6 (1), 34–67. Online available: <http://www.j-e-r-o.com/index.php/jero/article/view/407/192>.

Jan Retelsdorf. University of Hamburg, Von-Melle-Park 8, 20146 Hamburg, Germany. E-mail: jan.retelsdorf@uni-hamburg.de

Current themes of research:

Teacher's judgment bias. Science text comprehension. Motivation. Self-regulation.

Most relevant publications in the field of Psychology of Education:

Muntoni, F. & Retelsdorf, J. (2018). Gender-specific teacher expectations in reading - the role of teachers' gender stereotypes. *Contemporary Educational Psychology*, 54, 212–220. doi: 10.1016/j.cedpsych.2018.06.012

Retelsdorf, J., Köller, O., & Möller, J. (2011). On the effects of motivation on reading performance growth in secondary school. *Learning and Instruction*, 21, 550–559. doi: 10.1016/j.learninstruc.2010.11.001

Retelsdorf, J., Schwartz, K., & Asbrock, F. (2015). "Michael can't read!" – teachers' gender stereotypes and boys' reading self-concept. *Journal of Educational Psychology*, 107, 186–194. doi: 10.1037/a0037107