When a silent reading fluency test measures more than reading fluency: academic language features predict the test performance of students with a non-German home language

Karin Berendes¹ · Wolfgang Wagner¹ · Detmar Meurers² · Ulrich Trautwein¹

Published online: 7 July 2018
© The Author(s) 2018

Abstract

Silent reading is the primary mode of reading for proficient readers, and therefore, silent reading fluency is often assessed in research and practice. However, little is known about the validity of the tests administered to students with different language backgrounds. Given that academic language is assumed to be especially challenging for students with a non-German home language, one might wonder if effects of academic language proficiencies can be found on these tests. In the present study, we explored whether, owing to academic language demands, students with a non-German home language ($N=748$) would be found to be at a greater disadvantage than their monolingual-home peers ($N=1669$) on the most frequently used silent reading fluency test in Germany. Using differential item functioning (DIF) analyses, we found specific item difficulties to the disadvantage of the students with a non-German home language. This DIF was linked to the academic language features of the sentences.

Keywords Silent reading fluency measurement · nonmonolingual language background · Academic language · Test validity · Differential item functioning

Introduction

There are three general types of widely used reading tests: (a) decoding/accuracy tests, (b) reading comprehension tests such as those applied in the Programme for International Student Assessment (PISA; Organisation for Economic Co-operation

✉ Karin Berendes
karin.berendes@uni-tuebingen.de

¹ Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Europastraße 6, 72072 Tübingen, Germany

² Department of Linguistics, University of Tübingen, Tübingen, Germany

& Development—OECD, 2016), and (c) oral/silent reading fluency tests. Fluency tests measure “the ability to read a text quickly, accurately, and with proper expression” (National Institute of Child Health and Human Development [NICHD], 2000, Chapter 3, p. 5). Given that in silent reading, the prosodic indicators are measured through the reader’s “inner voice” (Féry, 2016), oral and silent reading fluency tests measure the same abilities to a large extent (Bar-Kochva, 2013; Share, 2008). The main difference between silent and oral reading is that speech production (voice articulation) speed is not measured by tests of silent reading fluency. Reading fluency tests are characterized by easy-to-understand content and language so that reading time is not affected by specific knowledge (e.g., cultural information), unknown vocabulary, or unfamiliar or complex syntactic constructions. The so-called error rate can be used as an indicator that reading speed and accuracy are really the main sources of variation. If the error rate is very low and does not vary across different groups of test-takers, the items can be expected to pose no problems for readers’ understanding. However, little information has been provided about the error rates of most of these tests, and not many studies have scrutinized the differential error rates across different groups of test-takers.

In this study, we sought to determine whether students with a non-German home language have problems understanding some items on reading fluency tests due to certain linguistic characteristics of the items. If this were the case, the test score would reflect not only basic reading skills but also certain linguistic competencies within this group. This would have two important implications: First, from a psychometric perspective, such a test would not be equally valid for students with monolingual versus nonmonolingual language backgrounds, and such a finding would have consequences for the use of this test in all diagnostic settings. Second, in terms of theory building, a closer look at the linguistic properties of the items used on reading fluency tests might aid the understanding of what makes reading difficult for nonmonolingual-home learners. In fact, if items that pose challenges to nonmonolingual-home learners are characterized by more cognitive academic language structures than other items, this would also support the need to differentiate between academic language and basic interpersonal language structures.

In our study, we applied a differential item functioning (DIF) approach to the most prominent German reading fluency test and studied the error rate in the silent reading fluency scores (sentence verification) of 11-year-old students with a non-German home language in comparison with their monolingual-home classmates. When a test item shows DIF, this indicates that equally proficient individuals from different groups do not have the same probability of answering this test item correctly (Angoff, 1993). In post hoc analyses, we used a range of academic language features as predictors of relative item-specific performance differences.

Silent reading fluency measures

Reading fluency is typically measured by the number of words or sentences that are processed correctly per a certain amount of time (e.g., Denton et al., 2011). A common approach for measuring silent reading fluency involves

sentence verification tasks such as on the sentence reading fluency subtest from the Woodcock-Johnson IV test (WJ IV; Schrank, Mather, & McGrew, 2014) and the Test of Sentence Reading Efficiency and Comprehension (OSREC; Wagner, Torgesen, Rashotte, & Pearson, 2010). These kinds of tasks require students to read declarative sentences silently and to indicate whether the statement is true or false (e.g., “All birds are blue”; “A cow is an animal”). A variation of this test procedure involves asking students to read simple interrogative sentences and to answer yes or no to each (e.g., “Is water dry?”; “Are some apples red?”; Kaufman Test of Educational Achievement III—KTEA III, silent reading fluency subtest; Kaufman & Kaufman, 2014). The test is supposed to capture the reader’s fluency in reading the input rather than the reader’s specific knowledge, linguistic competencies, or vocabulary knowledge. Therefore, the content and language of the sentences should be easy to understand, and the sentences should be based on fundamental world knowledge that young children are expected to know well. Moreover, the structure of the sentences should be simple, and the vocabulary that is included should be well-known. The test contains a verification task that is used to control for guessing. The test score is usually computed as the number of correct answers within a given period of time (usually 2–5 min).

In Germany, the most common test for measuring silent reading fluency (sentence verification) in secondary school is the Salzburger Lesescreening für die Klassenstufen 5–8 ([Salzburger Reading Screening for Grade Levels 5–8], SLS 5–8, Auer, Gruber, Mayringer, & Wimmer, 2005). In this screening, the lengths of the sentences range from 5 to 26 words, and the normative sample consists of monolingual German-speaking students only. According to the user manual, the error rates (error rate measured as the number of incorrect answers per student) for this group are low (varying between 0.5 and 0.6 sentences on average in Grades 5–8). Moreover, on average, in each grade level, fewer than 0.09 sentences are left out. Hence, the test seems to be successful in presenting “easy-to-understand” material. Given that the test is not constructed for non-native German-speaking students, the authors pointed out that students with a native language other than German could be at a disadvantage in this reading screening owing to lower language competencies and language comprehension problems. They also assessed data for students with a native language other than German and reported that these students correctly understood 5.5 fewer sentences than the monolingual German-speaking students whose performance ranged from $M=31.8$ ($SD=7.4$) to $M=41.6$ ($SD=9.2$) across Grades 5–8. However, the source of this quite substantial difference was not explained any further. It is possible that the students simply had lower basic reading skills and were therefore less fluent than their native German-speaking peers. However, it is also possible that these students had comprehension difficulties going back to certain language features (vocabulary, syntax) of the reading items. To consider this issue further, we first provide a brief overview of theories on native and non-native speakers’ sentence processing and then outline why academic language can be expected to be especially challenging for students with a non-German home language.

Native and non-native speakers' sentence processing

Different claims have been made concerning the differences in how native and non-native speakers process sentences. Some researchers have suggested that native and non-native speakers apply qualitatively different parsing routines when they process sentences. These researchers have argued that second-language learners tend to process structures shallowly (shallow structure hypothesis; Clahsen & Felser, 2006) or have trouble integrating different types of syntactic information (Felser, Roberts, Marinis, & Gross, 2003). By contrast, others have suggested that native and non-native speaker process sentences in fundamentally similar ways and have argued that differences result from an increased burden on capacity-limited cognitive resources (Hopp, 2010; McDonald, 2006). Still others have argued that the existing evidence indicates that differences in native and non-native speakers' sentence processing can best be explained in terms of different memory encoding, storage, and retrieval operations that promote successful language comprehension (e.g. Cunnings, 2017). They have suggested that non-native learners are more susceptible to interference during memory retrieval.

Academic language as a predictor of DIF on competence tests

In order to succeed across their educational careers, students must acquire basic interpersonal communication skills in general and cognitive academic language proficiency more specifically (Berendes, Dragon, Weinert, Heppt, & Stanat, 2013). Academic language, the language used to impart and acquire knowledge, is spoken in academic settings and used in school textbooks (for textbooks, see Berendes et al., 2018). It is designed to be precise and concise in order to refer to complex processes and to express complicated ideas in contextually reduced settings (see Cummins, 2008). Academic language proficiency develops through social interaction from birth but becomes differentiated from basic interpersonal communication skills after the early stages of schooling (Cummins, 2008). The acquisition of academic language proficiency is expected to be especially challenging for children who speak at least one home language that differs from the lingua franca of the society. In general, the quantitative input in the lingua franca for these children is lower given that they are also learning another language/ other languages (Unsworth, 2016). This is supposed to have only minor (or no) impact on their acquisition of basic interpersonal communication skills but more of an impact on their acquisition of cognitive academic language proficiencies in this language.

Academic language can be described by different descriptive, morpho-syntactic, and lexical features. For instance, on the descriptive level, sentence length is a good indicator of overall complexity. On the syntactic level, for example, a large number of noun phrases and a large number of nominalizations characterize academic language. On the lexical level, general academic vocabulary can be challenging. However, in the German language, syntactic features are assumed

to be more critical for comprehending academic language than lexical ones (e.g., Dehn, 2011; Prediger & Zindel, 2017).

During timed reading, the reader has only a limited or no opportunity to use compensatory mechanisms (Walczyk, 2000), and thus, nonmonolingual-home students' performance might suffer. However, until now, the role that academic language features play in reading fluency tests has not been studied.

Differential item functioning (DIF; de Boeck, 2008) analyses can be used to study whether reading fluency test items pose special comprehension difficulties for nonmonolingual-home students as compared with monolingual-home students. DIF analyses indicate that "equally able (or proficient) individuals, from different groups, do not have equal probabilities of answering the item correctly" (Angoff, 1993, p. 4). Thus, they measure the relative differences between two or more groups. First, one reference group and one or more focal groups are defined, and an item response theory (IRT) model is estimated for each group. The item parameters of each group are then compared with those in the reference group; (large) difference for some of the items indicate (substantial) DIF effects.

Several studies have investigated whether academic language features in test items are associated with larger comprehension difficulties for nonmonolingual-home students. Until now, however, all empirical evidence has stemmed from power tests rather than speed tests. A test is said to be a power test if it has "no rigid time limits: enough time is given for a majority of the test takers to complete all of the test items" (Landy & Conte, 2010, p. 127). Moreover, power-test items pose some difficulties, and mistakes are necessary to be able to differentiate between good and poor performances. By contrast, a test is said to be a speed test if "all items of the test are such that every subject who attempts the item answers it correctly and if there are so many items that no subject will be able to complete all items, within the time-limits specified for the test" (Jansen, 1997, p. 393). In a study of kindergarten children, differential effects were found for the processing of auditorily presented sentences to the disadvantage of children with a non-German home language. These effects could be explained by the number of prepositional phrases and the number of coordinations (Berendes, Wagner, Meurers, & Trautwein, 2015). Specific difficulties were also observed for elementary school children's reading performance (Heppt, Haag, Böhme, & Stanat, 2015). In this study (Heppt et al., 2015), DIF effects to the disadvantage of nonmonolingual students were explained by long and complex words, sentence length, and number of prepositional phrases. Haag, Heppt, Stanat, Kuhl, and Pant (2013) studied the linguistic complexity of German mathematics items and found that text length, general academic vocabulary, and number of noun phrases were unique predictors of DIF to the disadvantage of students who primarily heard or spoke a language other than German. Abedi, Leon, Wolf, and Farnsworth (2008) and Martiniello (2008, 2009) found that academic language features were the source of item bias for mathematics test items presented in English. In sum, previous research has found evidence that academic language demands can contribute to DIF on power tests. The present study is the first to investigate whether academic language demands are associated with larger comprehension difficulties for nonmonolingual-home students on a fluency test.

The present study

In the present study, we studied the silent reading fluency (sentence verification) error rate of nonmonolingual-home fifth graders in comparison with their monolingual-home classmates.

First, we investigated whether the reading fluency items were associated with specific comprehension difficulties (as measured by DIF) for the nonmonolingual-home students in comparison with their monolingual-home peers. Given that the time limit of the speed test did not allow for compensation strategies and that some items were rather long and could be expected to result in a high processing load, we expected to find DIF to the disadvantage of the nonmonolingual-home students (Hypothesis 1). DIF to the disadvantage of these students would indicate that the test measures not only reading fluency but other competencies as well. Thus, the test would not be equally valid for both groups.

Second, using post hoc analyses, we explored the role of academic language features in DIF. Given that academic language is expected to be especially challenging for children who speak at least one language at home other than the lingua franca of the society, and in accordance with findings from studies on power tests, we expected academic features to predict DIF to the disadvantage of the nonmonolingual-home students (Hypothesis 2). If this were the case, the test would also measure academic language competencies and not only reading fluency per se.

Method

Data collection and sample description

In this study, we used data from the TRAIN study (“Tradition und Innovation: Entwicklungsverläufe an Haupt- und Realschulen in Baden-Württemberg und Mittelschulen in Sachsen” [Tradition and innovation: Academic and psychosocial development in vocational track schools in the states of Baden-Württemberg and Sachsen]; Jonkmann, Rose, & Trautwein, 2013). The TRAIN study is a longitudinal study conducted in two German states (Baden-Württemberg and Sachsen) with a multistage sample of vocational track schools from which one to two classes are sampled per school (for details, see Rose, Jonkmann, Hübner, Sälzer, Lüdtke, & Nagy, 2013). The data used in the present study came from the first measurement point,¹ which took place in the 2008/2009 school year. All tests were administrated by trained test administrators. The sample comprised 2417 students (47% girls) from 105 schools and 130 classes in Grade 5 (age 11); 1669 of them were monolingual-home students (48% girls), and 748 students (45% girls) had at least one non-German home language.

¹ We also analyzed the test results of the students when they were older (Grade 6–8), but with increasing age, we found smaller differences on a descriptive level in the relative item difficulties between monolingual-home and nonmonolingual-home students.

Measures

Reading fluency

Reading fluency was measured with the Salzburger Lesescreening für die Klassenstufen 5–8 (SLS 5–8; Auer et al., 2005). The students were asked to read sentences that increased in length (5–26 words) and complexity. The content of the sentences was either reasonable (e.g., “Eine Woche hat sieben Tage” [“A week has 7 days”]) or not (e.g., “Die Sonne ist blau” [“The sun is blue”]). The students were asked to read the sentences quickly and to judge the veracity of the sentences (right or wrong). The test measures the number of sentences that a student can process correctly within the prescribed time of 3 min. There are two parallel versions (Forms A and B) that have both been applied in the TRAIN study. Each version contains 70 items with one overlapping item on the two forms (139 different items). The average number of items that were not reached was 38.81 ($Min=0$, $Max=67$).

Family language

We differentiated between two groups: Monolingual-home students and nonmonolingual-home students. Students who reported that they spoke only German at home were designated as monolingual-home students ($n=1669$). Students who reported speaking at least one language other than German at home ($n=748$) were designated as nonmonolingual-home students.

Linguistic analyses

To compute the linguistic complexity measures at the surface, morpho-syntactic, and lexical levels, we used a method proposed by Hancke, Vajjala, and Meurers (2012). For sentence segmentation, we employed OpenNLP (Apache Software Foundation, 2010) and integrated two types of syntactic analyses: the Stanford parser (Rafferty & Manning, 2008) for syntactic constituents and the MATE parser (Bohnet & Kuhn, 2012) for a dependency analysis integrating a morphological analysis. Moreover, we used the dlexDB (Heister et al., 2011) and the lexical semantic database GermaNet (Hamp & Feldweg, 1997) for lexical analyses.

The 14 features applied in this study were: sentence length, average word length, longest syntactic dependency, average number of dependencies per verb, average number of prepositional phrases per clause, number of noun phrases² per sentence, number of complex noun phrases per clause, lexical density, verb variation, modifier variation, verb-noun ratio, ratio of dative nouns to all nouns, ratio of compound nouns to all nouns, and ratio of auxiliary verbs to all verbs. The Appendix presents a description of each feature and our rationale for using it. Moreover, Table 3 in the Appendix presents the formulae used to calculate the scores on the features. Overall, our rationale for choosing these 14 features was the following: First, we picked the most common ones

² A noun phrase (NP) is “a phrase with a noun as its head” (Biber, Conrad, & Leech, 2002, p. 41).

from readability/complexity research, namely, sentence and word length. These two features have been used in traditional readability formulas for several decades now (see Benjamin, 2012) and are good indicators of syntactic and lexical complexity (for more information, see the [Appendix](#)). Second, we picked academic language features that are believed to pose special demands for nonmonolingual-home students. Of course, the features we considered represent only a small portion of all possible relevant linguistic features. However, we chose them such that they systematically made the academic language characteristics of the items explicit and showed sufficient variability.

Statistical analyses

In this study, two aspects of reading ability could be distinguished, namely, speed (i.e., the number of items completed) and precision (which is related to the proportion of correct solutions). The proportion of correct solutions was the target variable in our analyses and was modeled in an IRT framework. We focused on DIF in terms of item impact (i.e., construct-related group differences; Zumbo, 1999) regarding monolingual-home (MH) and nonmonolingual-home (NMH) students.

Both speed and precision can be assumed to depend on a common factor (mental speed; Roskam, 1997). Therefore, and in line with suggestions regarding missingness on items that were not reached in IRT models owing to time constraints (Pohl, Gräfe, & Rose, 2014; Rose, von Davier, & Nagengast, 2016), associations among the number of not-reached items and the latent variable (reflecting precision) were estimated in all models. The outcome in this study was comprised of the relative item difficulties (modeled in terms of item easiness) in the NMH group. These item difficulties were estimated in an IRT framework, which was based on a model proposed by de Boeck (2008): the cross-classified multilevel model with random effects for person and item. The difficulties were modeled as a latent variable at the item level (with item easiness represented by a distribution with mean zero and a freely estimated variance) in cross-classified multilevel models with categorical outcomes using a probit link function. At the student level, on the other hand, variability in precision across students was captured with another latent variable (assumed to be negatively correlated with the number of not-reached items as noted above). Further, group-specific thresholds representing the average item difficulty in each group were estimated. Therefore, in line with the DIF framework, relative (instead of absolute) difficulties were modeled at the item level (with mean zero in both groups as noted above).

Modeling the item difficulties as a latent variable instead of using point estimates has the same advantages as using a latent variable for precision at the student level: Biases due to unreliability can be avoided. Such biases play an increasing role when the measurement precision of the point estimates is small (i.e., their standard errors are large). With regard to item difficulties, this is particularly the case when the (sub-) sample size is rather small or the item difficulties are rather extreme (as in our study where all items were constructed so that almost all students were likely to solve them).

In a first model, at the student level, two latent variables representing precision in each group (MH, NMH) were established and regressed on the number of not-reached items (z-standardized within each group). The covariance between the

residuals of these latent variables was fixed to zero because it could not be estimated (i.e., each student belonged to only one of the groups per definition, so the coverage here was zero). The variances of the latent variables were calculated as the sum of the predicted variance and the residual variance of each latent variable.

At the item level, the item difficulties in the NMH group were regressed on the item difficulties in the MH group, represented by a second latent variable at the item level. In the ideal case of perfectly identical item difficulties in both groups, a regression weight of $b = 1.0$ would be expected, and the residual variance would be (close to) zero (i.e., R^2 would be close to 1.0). When R^2 is (significantly) smaller than 1.0, other potential predictors can be added to the regression to increase the amount of explained variance (e.g., linguistic features, as in this study).

In this study, a separate model was estimated for each feature. All models were estimated in Mplus (versions 7.3.1–8; Muthén & Muthén, 1998–2012) with the Bayes estimator.

Results

In a first step, we investigated group differences in the test scores, percent correct scores, error rates, and missing responses. With regard to the full item set (70 items for each test form), the test scores (sum of correctly answered items per test) were slightly higher in the MH group than in the NMH group (29.10 vs. 27.68, $p < .001$). Also, the percent correct scores were slightly higher in the MH group than in the NMH group (95.49 vs. 93.19%, $p < .001$). Further, the error rate (average number of incorrect responses) for the NMH students was statistically significantly higher than for the MH students (2.30 vs. 1.51, $p < .001$). On average, the last 39.18 (NMH) or 38.64 (MH) items were not reached ($p = .216$). However, the number of items that were skipped was small and comparable in size in the two groups (NMH: 0.80, MG: 0.73, $p = .683$). Also, the group difference for the sum of the not-reached and skipped items was not statistically significant ($p = .116$). In summary, the results showed that the test score advantage that the MH students showed over the NMH students could not be attributed to the MH students responding to a larger number of items. Rather, the MH students exhibited a smaller number of incorrect responses.

In the total sample and across both test forms, the percentages of correct versus incorrect solutions at the item level ranged from 41.67 to 99.09% ($M = 82.04\%$; $Mdn = 88.61\%$) with the number of valid (i.e., nonmissing) item responses ranging from 9 to 1321 ($M = 525.09$, $Mdn = 276.50$). Because there were extremely high percentages of correct solutions ($> 97.5\%$ in our sample), 13 items were excluded from further IRT analyses (Form A: six items; Form B: seven items).

To address Hypothesis 1, we estimated a DIF model that included the remaining test items and the number of not-reached items (which were included because they could be correlated with the latent variables at the person level, reflecting precision). The variances of the latent variables representing the item difficulties for precision in the NMH and MH groups were $\text{Var}(\Theta_{\text{NMH}}) = 0.30$ ($SD_{\text{NMH}} = 0.55$) and $\text{Var}(\Theta_{\text{MH}}) = 0.35$ ($SD_{\text{MH}} = 0.59$). The thresholds—reflecting the location of the average item difficulty—in the NMH and MH groups were located at -1.40 and -1.72

on the probit scale, respectively. This corresponds to a probability of solving an item of average difficulty for a student with average precision in each group of 92.95% (NMH) and 95.75% (MH). These results closely matched the “simple” group means of the percent correct scores based on the set of selected items (with items coded 1 = correct, 0 = incorrect or missing because of nonresponse or because of the multi-matrix design with two different test forms). These were 91.92% (NMH) and 94.72% (MH), and the difference between the groups was statistically significant ($p < .001$).

At the item level, the variance of the item difficulties in the MH group was $\text{Var}(\beta_{\text{MH}}) = 0.13$ ($SD = 0.35$). These item difficulties predicted the item difficulties in the NMH group with a regression weight of $b = 1.15$ with the 95% confidence limits [0.98, 1.31] including 1.0. In line with our expectations (Hypothesis 1), the prediction was not perfect (as would be expected in the absence of DIF between the groups) with $R^2 = .90$. Therefore, linguistic features were added to explain differences in item difficulties between groups (Hypothesis 2). Because there were some strong associations between the features ranging from $-.58 \leq r \leq .80$ ($Mdn: 0.02$; see Table 1, which also reports the means and standard deviations of each feature), we refrained from estimating a single model with all features simultaneously but instead estimated separate models for each feature.

The inclusion of the different features in the model showed statistically significant results in three of the 14 cases (Table 2): As expected, relatively higher difficulties were found for items consisting of more words for the NMH students ($b = 0.01$, $p = .02$, one-tailed). For items with six more words, for example, the difficulty would be expected to be approximately 0.1 SD higher (pooled $SD = [(0.30 + 0.35)/2]^{0.5} = 0.57$; see above) for NMH students than for MH students. A descriptive comparison of the percentages of correct solutions across all items with a small number of words (below or equal to $Mdn = 12$, $k = 74$ items) showed only slightly higher percentages for the MH group ($M = 89.55\%$, $Mdn = 94.74$) than for the NMH group ($M = 84.21\%$, $Mdn = 90.34\%$). Items consisting of many words (above the median, $k = 53$ items with responses in both groups), on the other hand, showed much higher percentages of correct solutions for the MH group ($M = 72.21\%$, $Mdn = 71.43\%$) compared with the NMH group ($M = 62.66\%$, $Mdn = 69.23\%$).

In addition, statistically significant effects were found for the number of noun phrases per sentence ($b = 0.04$, $p < .01$, one-tailed) and the verb-noun ratio ($b = -0.06$, $p = .02$, one-tailed). Regarding percentages of correct answers for low and high (median split) numbers of noun phrases per sentence ($Mdn = 4$), a descriptive comparison showed larger group differences (MH, NMH) for items with more (MH: $M = 73.87\%$, $Mdn = 71.43\%$, NMH: $M = 62.55\%$, $Mdn = 66.67\%$, $k = 48$ items with responses in both groups) versus smaller numbers of noun phrases per sentence (MH: $M = 87.45\%$, $Mdn = 93.98\%$, NMH: $M = 82.91\%$, $Mdn = 89.24\%$, $k = 79$ items). With regard to the verb-noun ratio, smaller descriptive group differences (MH, NMH) were also found for items with a larger ($Mdn = 0.67$) verb-noun ratio (MH: $M = 78.00\%$, $Mdn = 80.00\%$, NMH: $M = 72.44\%$, $Mdn = 75.62\%$, $k = 63$ items with responses in both groups) compared with items with a lower verb-noun ratio (MH: $M = 86.57\%$, $Mdn = 92.79\%$, NMH: $M = 77.96\%$, $Mdn = 85.63\%$, $k = 64$ items). Thus, a larger number of noun phrases per sentence and a larger number of nouns in comparison with verbs contributed to specific disadvantages for nonmonolingual-home students.

Table 1 Intercorrelations of the 14 linguistic features

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1 Sentence length (in words)	1													
2 Ø Word length (in syllables)	-.52	1												
3 Longest syntactic dependency (in words)	.78	-.38	1											
4 Ø Number of dependencies per verb	.38	-.12	.23	1										
5 Ø Number of prepositional phrases per clause	-.09	.14	-.10	.19	1									
6 Number of noun phrases per sentence	.80	-.43	.66	.31	-.05	1								
7 Ø Number of complex noun phrases per clause	.10	.02	.13	.03	.15	.17	1							
8 Lexical density	-.40	.38	-.30	-.28	.11	-.31	.22	1						
9 Verb variation	.12	-.20	.14	-.03	-.07	.05	-.27	-.18	1					
10 Modifier variation	.21	-.01	.08	.17	-.14	-.12	.19	.25	-.29	1				
11 Verb–noun ratio	.09	-.21	.04	-.15	-.29	-.16	-.28	-.18	.70	.17	1			
12 Ratio of dative nouns to all nouns	-.11	.12	-.15	.08	.60	-.18	.03	.06	.26	-.02	.15	1		
13 Ratio of compound nouns to all nouns	-.26	.35	-.19	.05	.05	-.32	-.04	.00	.03	.17	.11	.16	1	
14 Ratio of auxiliary verbs to all verbs	-.06	.05	-.05	-.27	-.31	-.13	.07	-.20	-.58	.24	-.10	-.33	.00	1
Mean	11.70	2.01	8.37	2.82	0.83	4.12	0.48	0.54	0.29	0.18	0.91	0.34	0.30	0.25
Standard deviation	4.31	0.38	4.34	0.73	0.69	1.47	0.60	0.10	0.16	0.16	0.76	0.29	0.31	0.31

Statistically significant correlations ($p \leq .05$, two-tailed) are presented in bold

Table 2 Predicting item difficulty with the 14 linguistic features: unstandardized regression weights and *p* values

	Linguistic features	<i>b</i>	<i>p</i>
1	Sentence length (in words)	0.01	.02
2	Ø Word length (in syllables)	−0.01	.47
3	Longest syntactic dependency (in words)	0.01	.12
4	Ø Number of dependencies per verb	0.01	.43
5	Ø Number of prepositional phrases per clause	−0.01	.41
6	Number of noun phrases per sentence	0.04	.00
7	Ø Number of complex noun phrases per clause	0.08	.06
8	Lexical density	0.28	.18
9	Verb variation	−0.16	.17
10	Modifier variation	0.06	.34
11	Verb–noun ratio	−0.06	.02
12	Ratio of dative nouns to all nouns	−0.11	.10
13	Ratio of compound nouns to all nouns	−0.03	.32
14	Ratio of auxiliary verbs to all verbs	−0.07	.18

Regression weights were multiplied by -1 to allow for an interpretation in terms of item difficulty. Values larger than zero indicate relatively higher item difficulties in the NMH group with increasing values of the respective feature. The reported *p* values refer to one-tailed tests

Discussion

In this study, we explored whether 11-year-old nonmonolingual-home students would suffer from comprehension difficulties that would leave them at a disadvantage in comparison with monolingual-home students on a reading fluency test. We indeed found DIF to the disadvantage of the nonmonolingual-home students. The DIF could be explained by three academic language features of the sentences: sentence length, number of noun phrases per sentence, and verb–noun ratio.

The findings suggest that the nonmonolingual-home students had specific comprehension difficulties on the reading fluency test. More specifically, the measurement invariance assumption was violated for the precision component of the test (which impacts the test score through the number of incorrect solutions). Therefore, for the comparison of students with a different language background, the selection of specific items with respect to features of the sentences has an impact on the results. It is likely that these comprehension difficulties can be attributed to an inappropriate adjustment of reading speed (inappropriate *strategic fluency*; see Topping, 2006) when students read sentences with linguistic structures that pose a special challenge for them. A fast, heuristic, superficial pseudo-parsing strategy thus results in incomplete, underspecified, or shallow representations and subsequently often in a wrong answer (see Christianson, 2016; Karimi & Ferreira, 2016).

What are the specific academic language features that caused the DIF in our study? First, feature sentence length explained the DIF that disadvantaged the nonmonolingual-home students. This finding is consistent with results from other

studies that employed power tests (e.g., English language: Abedi, Leon, Wolf, & Farnsworth, 2008; German language: Berendes et al., 2015; Haag et al., 2013; Heppt et al., 2015). Berendes et al. (2015) found significant group differences in item difficulties on a grammar comprehension test for 5-year-old children who spoke German at home versus children who spoke a language other than German at home. As expected, some of the item-specific difficulties could be explained by the increased sentence length. Heppt et al.'s (2015) results also indicated that longer sentences posed a special burden for nonmonolingual-home fourth graders. Their results showed DIF that worked against nonmonolingual-home students on reading comprehension items. This could be explained by average sentence length as well as some other features. Haag et al. (2013) found comparable results for text length. The research group used data from a state-wide mathematics test for German third graders and found that text length was a unique contributor to DIF that worked against second language (L2) learners. Thus, the length of the input stimuli seems to be a good predictor of DIF that worked to the disadvantage of nonmonolingual-home students independent of the age of the sample (kindergarten, elementary school, secondary school), the kind of language skill under study (listening, reading), or the test used (power test, speed test). This finding is plausible because sentence length may in general be assumed to be positively associated with syntactic complexity.

Second, the number of noun phrases per sentence was relevant in our study and has also been found to be a predictor of DIF in some studies using power tests (Haag et al., 2013) but not in others (Berendes et al., 2015; Heppt et al., 2015). Different results could probably be attributed to sample characteristics (e.g., language and socioeconomic background of the students), the grain size of the stimulus material (e.g., sentence level, text level), purpose of the stimulus material and thus construction properties (e.g., test items, items from a textbook), the language skills under study (e.g., reading comprehension, listening comprehension), or the statistical method that was applied (e.g., DIF, interactions). To study this issue further, future studies could systematically vary the kind of stimulus material and the number of noun phrases in the items under study.

Third, the verb-noun ratio also affected DIF in our study. To our knowledge, this feature has not been assessed in other studies with (German) samples before. Thus, unfortunately, we cannot compare our findings with results from other studies. However, the findings are in line with the theoretical consideration that academic language structures are especially challenging for nonmonolingual-home students because a larger number of nouns than verbs is a typical indicator of this language register.

One feature, prepositional phrase, which has been a good predictor of DIF in other studies with children (kindergarten: Berendes et al., 2015; fourth graders: Heppt et al., 2015), did not contribute to explaining the DIF effects in our study. The different results may be attributable to age differences. In the studies by Berendes et al. (2015) and Heppt et al. (2015), the children were younger than the children in the present study. Prepositional phrases cannot be expected to pose a special hurdle for L2 students in secondary school as they do for younger L2 learners (see also Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006).

In sum, we can conclude that the lower academic language proficiencies of the nonmonolingual-home students in our study caused them to suffer from

disadvantages in their reading fluency outcomes. These disadvantages could be attributed to longer sentences, a large number of noun phrases, and a rather nominal style. The results suggest that the fluency test we used measures different kinds of fluency, namely, reading fluency for simple sentences versus reading fluency for complex sentences. This finding may hold for other fluency tests too.

Whereas the documented DIF effects may constitute validity issues for existing tests, they also point to opportunities for research and practice. More specifically, we suggest that reading fluency tests should be systematically divided into two subtests with different norms: one subtest with items written in simple language requiring only basic interpersonal communication skills and one subtest with more sophisticated and complex language requiring a proficiency in academic language. In addition, when a separate subtest with sentences using academic language is constructed, certain academic language features can be systematically included. Furthermore, tablets and eye-tracking can be applied to measure the time spent on each sentence. A computer-based administration capturing both item response accuracy and response times would allow researchers to use conditional item response theory such as illustrated by Petscher, Mitchell, and Foorman (2015). This would shed more light on the demands that certain linguistic structures pose for (nonmonolingual-home) students, thus providing information that will be relevant for both research and practice.

As we applied DIF analyses to a large data set, the present study provides some important results with practical implications about group differences between monolingual-home and nonmonolingual-home students on a German reading fluency test. However, some limitations need to be mentioned. For instance, we did not control for working memory capacity, which can be expected to be an important factor in the ability to process syntactically complex sentences (King & Just, 1991). In addition, it was not possible to differentiate between students with different non-German home languages in our data set because of group sizes. In order to study the effects of a specific language background, other different language backgrounds as well as the start, duration, and intensity of contact with the German language could be considered in future studies. Moreover, in many cases, the occurrence of the academic language features was correlated with the item position (e.g., shorter sentences at the beginning of the test, longer sentences at the end of the test). Therefore, in future studies, item position could be assigned randomly. In addition, although our sample was representative of 11-year old children from different regions (e.g., rural, urban, different states) in two States of Germany, the results cannot be generalized to other (age) groups. Given that language develops continuously with daily use, certain linguistic features cannot be expected to remain particularly demanding for a long period of time (see also Footnote 1). Thus, in other age groups, other linguistic features are likely to pose a special burden for nonmonolingual-home students.

The results of the current study have important implications for reading fluency assessment. Unfortunately, reading fluency is usually considered a lower level reading skill that should be mastered early in literacy acquisition (Rasinski, 2012). Building on this, many fluency tests measure only the WCPM (i.e., words correct per minute) or fluency for short and simple sentences. However, in secondary school, silent reading of sentences and texts that are written in academic language is usually

required, and the source of reading problems is still a lack of fluency for many students (Rasinski, 2012). Therefore, reading fluency tests for students in secondary school should focus on academic language or should contain an academic language subtest. Until now, to the best of our knowledge, there have been no tests for assessing silent reading fluency with items that are explicitly constructed with academic language. Thus, further research and development is needed in this direction.

Our findings also have practical implications for the language support and teaching of nonmonolingual-home students. With regard to language support, our findings clearly indicate that specific academic language features should be the focus of instruction in order to help these students catch up with their monolingual-home peers. Regarding teaching, teachers should be aware of the potential difficulties that specific academic language features might pose to nonmonolingual-home students and should try to address them during lessons.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix

See Table 3.

Table 3 Calculation formulas for the 14 linguistic features

	Feature name	Calculation formula
1	Sentence length (in words)	Number of words per sentence
2	Ø Word length (in syllables)	Number of syllables/number of words
3	Longest syntactic dependency (in words)	Max(length(dependency))
4	Ø Number of dependencies per verb	Number of dependents of verbs/number of verbs with dependents
5	Ø Number of prepositional phrases per clause	Number of prepositional phrases/number of clauses
6	Number of noun phrases per sentence	Number of noun phrases per sentence
7	Ø Number of complex noun phrases per clause	Number of complex nominal phrases/number of clauses
8	Lexical density	Number of lexical tokens/number of tokens
9	Verb variation	Number of verb-types/number of lexical words
10	Modifier variation	(Number of adverb + adjective types)/number of lexical tokens
11	Verb–noun ratio	Number of verbs/number of nouns
12	Ratio of dative nouns to all nouns	Number of nouns with dative case/number of nouns
13	Ratio of compound nouns to all nouns	Number of compound nouns/number of nouns
14	Ratio of auxiliary verbs to all verbs	Number of auxiliary verbs/number of verbs

Descriptions of the 14 linguistic features

Sentence length (Feature 1) Sentence length was measured as the number of words per sentence. This measure is a good surface indicator of syntactic complexity and pertains to the best indicators of text readability (Nickel, 2011). Furthermore, an increase in the number of academic language structures goes hand in hand with an increase in sentence length (Heppt, Dragon, Berendes, Stanat, & Weinert, 2012). In general, longer sentences create a higher load on working memory and are harder to understand than shorter ones (Bamberger & Vanecek, 1984).

Average word length (Feature 2) This measure provides a count of the average number of syllables per word in each sentence. It is one of the most commonly used measures of lexical complexity in traditional readability research. Lenzner (2014) explained that “word length has a direct effect on the ease with which a text can be read: The longer a word is, the more difficult it is to comprehend” (p. 681).

Longest syntactic dependency (Feature 3) This feature refers to the longest distance between a word and its dependent in a sentence. The feature reflects the central idea of Gibson’s Dependency Locality Theory (DLT), which is that “the cost of integrating two elements (such as a head and a dependent [...]) depends on the distance between the two” (Gibson, 2000, p. 95–96). Thus, longer distances should pose greater processing demands than shorter ones (Temperley, 2007).

Average number of dependencies per verb (Feature 4) This feature is based on the dependence-valence theory by Tesnière (1980) and provides a count of the number of actants/dependencies per verb. The more dependencies a verb has, the harder it is to process these structures.

Average number of prepositional phrases per clause (Feature 5) A prepositional phrase consists of a preposition followed by a noun phrase. Prepositional phrases “contribute to the length and complexity of sentences” (Butler, Lord, Stevens, Borrego, & Bailey, 2004, p. 64) and are a well-known hurdle in language acquisition (Turgay, 2010). Moreover, prepositions are often ignored by readers (Jorgensen, 2011).

Number of noun phrases (see Footnote 2) per sentence (Feature 6) Large numbers of noun phrases (NPs) are found more often in academic language than in everyday language (e.g., Bailey, Butler, Stevens, & Lord, 2007), and large numbers of NPs are expected to pose a special burden for L2 students because “the integration cost increases with the number of new discourse referents that are introduced between the phrasal heads that must be integrated” (Gordon, Hendrick, & Johnson, 2004, p. 98). For example, to understand the nonsense sentence: “Etwas Wolliges, Warmes, das man meistens im Winter aufsetzt, wenn es schneit und wenn ein kalter Wind weht, ist eine Pfütze” [“A woolly and warm thing that one usually puts on in winter when it is snowing and when a cold wind is blowing is a puddle”], the interactions between the different noun phrases must be understood. Also, the processing cost

required to integrate the structures in the sentence is high given the distance over which the integration occurs.

Number of complex noun phrases per clause (Feature 7) Complex noun phrases occur more often in academic writing and indicate syntactic complexity (Crossley & McNamara, 2014). Moreover, such phrases pose a special burden to L2 learners (Gürsoy, 2010). Complex nominals are defined as comprised of one of the following three conditions (Cooper, 1976): (a) nouns with an adjective, possessive, prepositional phrase, relative clause, participle, or appositive, (b) nominal clauses, or (c) gerunds and infinitives in the subject position. The occurrences of these three conditions were calculated by counting the occurrences of the respective patterns in the syntactic parse tree. A clause is defined as a syntactic structure consisting of a subject and a finite verb. This feature is important to consider when studying the reading competencies of students because complex noun phrases use various demanding syntactic structures and therefore pose a considerable challenge to less experienced readers (Schmidt, 1993).

Lexical density (Feature 8) This feature traces back to Ure (1971) and measures the relationship between function and content words and is an indicator of academic language. Academic language is believed to contain a larger number of content words, whereas the basic interpersonal language register is believed to have more grammatical words. This reflects the fact that more information is packed into each sentence in academic language.

Verb variation (Feature 9) Verb variation measures the ratio of unique verbs to the total number of lexical tokens. It is a measure of lexical variation that is closely related to lexical diversity measures, such as type token ratios. Unlike lexical diversity, lexical variation is restricted to certain word categories, which makes it possible to assess the sophistication of this specific category. Verb variation is an especially interesting measure for our purposes because verbs generally make large contributions to the meaning of a sentence, and they are frequent enough to support reliable measures even when the overall amount of data is limited.

Modifier variation (Feature 10) Modifier variation refers to the ratio of the total number of unique adjectives and adverbs in a text to the total number of lexical words. Adjectives and adverbs are typical modifiers. Metaphorically speaking, they are embellishing ornaments that contribute to the linguistic elaboration of nominal and verbal structures. They are not necessary and are not as predictable as many other constituents of a sentence. To build a complete sentence, a lexical verb is needed along with one or more constituents that satisfy the requirements of that particular verb. Besides these obligatory constituents (arguments), a sentence often contains optional elements (modifiers). From the psycholinguistic literature on ambiguity resolution, it is well-known that the human sentence parser finds it easier to process arguments than modifiers (e.g., Clifton, Speer, & Abney, 1991). We expected this to hold in nonambiguous contexts as well.

Verb–noun ratio (Feature 11) The verb-noun ratio is measured by dividing the number of verbs by the number of nouns. This feature is theoretically based on the fact that nouns occur more frequently in written language, and texts with a larger number of nouns than verbs have a high density of information, are therefore more demanding, and are harder to process (Heimann Mühlenbock, 2013). Nouns and verbs are complementary in the sense that a writer often has a choice about whether to express something in a verbalized or nominalized form. Texts with a low verb-noun ratio typically contain more nominalizations, and “nominalization allows an extended explanation to be condensed into a complex noun phrase” (Schlepppegrell, 2001, p. 443). Therefore, students have to process more ideas per clause when reading texts with nominalizations, and some students may have difficulties constructing the underlying meaning (Fang, Schlepppegrell, & Cox, 2006).

Ratio of dative nouns to all nouns (Feature 12) This feature measures the ratio of the number of nouns with dative case markers (indirect objects) to all nouns in a text. The dative, especially the plural dative, is very rarely used in oral language (see Bast, 2003; Turgay, 2011) and therefore takes longer to acquire. Different studies have shown that children with non-German home languages have great difficulties with dative nouns (for an overview, see Turgay, 2011). Moreover, learners sometimes confuse “der” for dative singular feminine with “der” for nominative singular masculine (Kaltenbacher & Klages, 2007) and confound dative and accusative. Therefore, it can be demanding for them to distinguish the meaning of sentences such as “Ich gebe dem Jungen den Hund” [“I give the dog to the boy”] versus “Ich gebe den Jungen dem Hund” [“I give the boy to the dog”].

Ratio of compound nouns to all nouns (Feature 13) Compound nouns are complex cognitive representations (Libben & Jarema, 2006) and can cause L2 learners to make errors, especially if their home language has a different compound structure: “Bilinguals whose L1 has, for example, right-headed compounds and whose L2 has left-headed compounds, might be expected to make errors of constituent misordering and produce right-headed L2 compounds” (Levy, Goral, & Obler, 2006, p. 130).

Ratio of auxiliary verbs to all verbs (Feature 14) This feature is a measure of the ratio of auxiliary verbs to all verbs. Linguistic constructions with auxiliary verbs may be more difficult in the German language because of the distance between the auxiliary and the main verb (e.g., “Einige Astronauten sind schon mit einer Rakete zum Mond geflogen” [“Some astronauts have already flown to the moon in a rocket”]).

References

- Abedi, J., Leon, S., Wolf, M. K., & Farnsworth, T. (2008). Detecting test items differentially impacting the performance of ELL students. In M. K. Wolf, J. L. Herman, J. Kim, J. Abedi, S. Leon, & N. Griffin, et al. (Eds.), *Providing validity evidence to improve the assessment of English language learners* (pp. 55–75). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

- Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale, NJ: Lawrence Erlbaum.
- Apache Software Foundation. (2010). *OpenNLP*. Retrieved February 13, 2016, from <http://opennlp.apache.org>.
- Auer, M., Gruber, G., Mayringer, H., & Wimmer, H. (2005). *SLS 5–8. Salzburger Lese-Screening für die Klassenstufen 5–8 [Salzburger reading screening for grade levels 5–8]*. Bern: Huber.
- Bailey, A. L., Butler, F. A., Stevens, R., & Lord, C. (2007). Further specifying the language demands of school. In A. L. Bailey (Ed.), *The language demands of school: Putting academic English to the test* (pp. 103–156). New Haven, CT: Yale University Press.
- Bamberger, R., & Vanecek, E. (1984). *Lesen – Verstehen – Lernen – Schreiben. Die Schwierigkeitsstufen von Texten in deutscher Sprache [Reading—Comprehension—learning—writing. The levels of difficulty of texts in the German language]*. Wien: Jugend und Volk. <https://doi.org/10.2307/3530491>.
- Bar-Kochva, I. (2013). What are the underlying skills of silent reading acquisition? A developmental study from kindergarten to the 2nd grade. *Reading and Writing: An Interdisciplinary Journal*, 26, 1417–1436.
- Bast, C. (2003). *Der Altersfaktor im Zweitspracherwerb – die Entwicklung der grammatischen Kategorien Numerus, Genus und Kasus in der Nominalphrase im ungesteuerten Zweitspracherwerb des Deutschen bei russischen Lernerinnen*. [The age factor in second language acquisition—The development of number, gender, and case in a nominal phrase in uncontrolled second language acquisition of German in Russian learners]. Dissertation, University of Cologne.
- Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24, 63–88. <https://doi.org/10.1007/s10648-011-9181-8>.
- Berendes, K., Dragon, N., Weinert, S., Heppt, B. & Stanat, P. (2013). Hürde Bildungssprache? Eine Annäherung an das Konzept Bildungssprache und aktuelle empirische Forschungsergebnisse [The obstacle academic language? A convergence towards the concept of “academic language” on the basis of current empirical research]. In A. Redder & S. Weinert (Hrsg.), *Sprachförderung und Sprachdiagnostik. Perspektiven aus Psychologie, Sprachwissenschaft und empirischer Bildungsforschung* (S. 17–41). Münster: Waxmann.
- Berendes, K., Vajjala, S., Meurers, D., Bryant, D., Wagner, W., Chinkina, M., et al. (2018). Reading demands in secondary school: Does the linguistic complexity of textbooks increase with grade level and the academic orientation of the school track? *Journal of Educational Psychology*, 110(4), 518–543.
- Berendes, K., Wagner, W., Meurers, D., & Trautwein, U. (2015). Grammatikverständnis von Kindern unterschiedlicher sprachlicher und sozioökonomischer Herkunft [Grammar comprehension of children with different language and socioeconomic backgrounds]. *Frühe Bildung*, 4(3), 126–134.
- Biber, D., Conrad, S., & Leech, G. (2002). *Longman student grammar of spoken and written English*. Harlow, Essex: Pearson Education Limited.
- Bohnet, B., & Kuhn, J. (2012). The best of both worlds—A graph-based completion model for transition-based parsers. In *Proceedings of the 13th conference of the European chapter of the Association for Computational Linguistics (EACL)* (pp. 77–87). Retrieved from <http://dl.acm.org/citation.cfm?id=2380828>.
- Butler, F. A., Lord, C., Stevens, R., Borrego, M. & Bailey, A. L. (2004). *An approach to operationalizing academic language for language test development purposes: Evidence from fifth-grade science and math* (CSE Report 626). Los Angeles, CA: University of California.
- Christianson, K. (2016). When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing. *The Quarterly Journal of Experimental Psychology*, 69(5), 817–828. <https://doi.org/10.1080/17470218.2015.1134603>.
- Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics*, 27, 3–42.
- Clifton, C., Speer, S., & Abney, S. P. (1991). Parsing arguments: Phrase structure and argument structure as determinants of initial parsing decisions. *Journal of Memory and Language*, 30, 251–271.
- Cooper, T. C. (1976). Measuring written syntactic patterns of second language learners of German. *The Journal of Educational Research*, 69(5), 176–183. <https://doi.org/10.1080/00220671.1976.10884868>.

- Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, 26, 66–79. <https://doi.org/10.1016/j.jslw.2014.09.006>.
- Cummins, J. (2008). BICS and CALP: Empirical and theoretical status of the distinction. In B. Street & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (Vol. 2, pp. 71–83)., Literacy New York, NY: Springer.
- Cummings, I. (2017). Parsing and working memory in bilingual sentence processing. *Bilingualism: Language and Cognition*, 20(4), 659–678. <https://doi.org/10.1017/s1366728916000675>.
- de Boeck, P. (2008). Random item IRT models. *Psychometrika*, 73(4), 533–559. <https://doi.org/10.1007/s11336-008-9092-x>.
- Dehn, M. (2011). Elementare Schriftkultur und Bildungssprache [Elementary written culture and academic language]. In S. Fürstenau & M. Gomolla (Eds.), *Migration und schulischer Wandel: Mehrsprachigkeit* (pp. 129–151). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Denton, C. A., Barth, A. E., Fletcher, J. M., Wexler, J., Vaughn, S., Cirino, P. T., et al. (2011). The relations among oral and silent reading fluency comprehension in middle school: Implications for identification and instruction of students with reading difficulties. *Scientific Studies of Reading*, 15(2), 109–135.
- Fang, Z., Schleppegrell, M. J., & Cox, B. E. (2006). Understanding the language demands of schooling: Nouns in academic registers. *Journal of Literacy Research*, 38(3), 247–273.
- Felser, C., Roberts, L., Marinis, T., & Gross, R. (2003). The processing of ambiguous sentences by first and second language learners of English. *Applied Psycholinguistics*, 24, 453–489.
- Féry, C. (2016). *Intonation and prosodic structure*. Cambridge: Cambridge University Press.
- Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain* (pp. 95–126). Cambridge, MA: MIT Press. Retrieved from http://www2.bcs.rochester.edu/sites/jaeager/teaching/LabSyntax2006/readings/Gibson_2000.pdf.
- Gordon, P. C., Hendrick, R., & Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *Journal of Memory and Language*, 51, 97–114.
- Gürsoy, E. (2010). Komplexe Nominalgruppen im Deutschen und in weiteren Sprachen [Complex nominal groups in German language and in other languages]. In: *Kompetenzzentrum ProDaZ*. Retrieved from http://www.uni-due.de/imperia/md/content/prodaz/komplexe_nominalgruppen.pdf.
- Haag, N., Heppt, 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*, 28, 24–34.
- Hamp, B. & Feldweg, H. (1997). GermaNet—A lexical-semantic net for German. In P. Vossen, N. Calzolari, G. Adriaens, A. Sanfilippo, & Y. Wilks (Eds.), *Proceedings of the ACL/EACL-97 workshop on automatic information extraction and building of lexical semantic resources for NLP applications* (pp. 9–15), Madrid.
- Hancke, J., Vajjala, S., & Meurers, D. (2012). Readability classification for German using lexical, syntactic and morphological features. In *Proceedings of the 24th international conference on computational linguistics (COLING)* (pp. 1063–1080). Retrieved from <http://aclweb.org/anthology/C12-1065.pdf>.
- Heimann Mühlenbock, K. (2013). *I see what you mean: Assessing readability for specific target groups*. *Data linguistica* (Vol. 24). Gothenburg: University of Gothenburg.
- Heister, J., Würzner, K., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., et al. (2011). dlexDB – eine lexikalische Datenbank für die psychologische und linguistische Forschung [dlexDBA—Lexical database for the psychological and linguistic research]. *Psychologische Rundschau*, 62, 10–20. <https://doi.org/10.1026/0033-3042/a000029>.
- Heppt, B., Dragon, N., Berendes, K., Stanat, P., & Weinert, S. (2012). Beherrschung von Bildungssprache bei Kindern im Grundschulalter [Mastery of academic language of children in elementary school]. *Diskurs Kindheits- und Jugendforschung*, 3, 349–356.
- Heppt, 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.
- Hopp, H. (2010). Ultimate attainment in L2 inflectional morphology: Performance similarities between non-native and native speakers. *Lingua*, 120, 901–931.
- Jansen, M. G. H. (1997). Rasch's model for reading speed with manifest explanatory variables. *Psychometrika*, 62(3), 393–409.

- Jonkmann, K., Rose, N., & U. Trautwein (Hrsg.) (2013). *Tradition und Innovation: Entwicklungsverläufe an Haupt- und Realschulen in Baden-Württemberg und Mittelschulen in Sachsen – Abschlussbericht für die Länder Baden-Württemberg und Sachsen* [Tradition and innovation: Academic and psychosocial development in vocational track schools in the states of Baden-Württemberg and Sachsen]. Tübingen: Projektbericht an die Kultusministerien der Länder.
- Jorgensen, R. (2011). Language, culture and learning mathematics: A Bourdieuan analysis of Indigenous learning. In C. Wyatt-Smith, et al. (Eds.), *Multiple perspectives on difficulties in learning literacy and numeracy* (pp. 315–329). Dordrecht: Springer.
- Kaltenbacher, E., & Klages, H. (2007). Deutsch für den Schulstart: Zielsetzungen und Aufbau eines Förderprogramms [Competent in German language at the beginning of school: Aims and development of a support program]. In B. Ahrenholz (Ed.), *Deutsch als Zweitsprache – Voraussetzungen und Konzepte für die Förderung von Kindern und Jugendlichen mit Migrationshintergrund* (pp. 135–154). Freiburg: Fillibach.
- Karimi, H., & Ferreira, F. (2016). Good-enough linguistic representations and online cognitive equilibrium in language processing. *The Quarterly Journal of Experimental Psychology*, 69(5), 1013–1040.
- Kaufman, A. S., & Kaufman, N. L. (2014). *Kaufman test of educational achievement* (3rd edn.). Bloomington, MN: NCS Pearson.
- King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30(5), 580–602.
- Landy, F. J., & Conte, J. M. (2010). *Work in the 21st century: An introduction to industrial and organizational psychology* (3rd ed.). Malden, MA: Wiley-Blackwell.
- Lenzner, T. (2014). Are readability formulas valid tools for assessing survey question difficulty? *Sociological Methods & Research*, 43(4), 677–698. <https://doi.org/10.1177/0049124113513436>.
- Levy, E. S., Goral, M., & Obler, L. K. (2006). Doghouse/Chien-maison/Niche: Approaches to the understanding of compound processing in bilinguals. In G. Libben & G. Jarema (Eds.), *The representation and processing of compound words* (pp. 125–144). Oxford: Oxford University Press.
- Libben, G., & Jarema, G. (2006). *The representation and processing of compound words*. Oxford: Oxford University Press.
- Martiniello, M. (2008). Language and the performance of English language learners in math word problems. *Harvard Educational Review*, 78(2), 333–368.
- Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment*, 14(3/4), 160–179. <https://doi.org/10.1080/10627190903422906>.
- McDonald, J. (2006). Beyond the critical period: Processing based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language*, 55, 381–401.
- Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- Nickel, S. (2011). Textschwierigkeit objektivieren: Der Lesbarkeitsindex LIX. Wie schwierig sind Lesetexte in der Alphabetisierung? [Objectifying text difficulty: The readability index LIX. How difficult are reading texts in alphabetization?]. *Alfa-Forum*, 76, 30–32.
- Organisation for Economic Co-operation and Development—OECD. (2016). *PISA 2015 results (Volume I): Excellence and equity in education*. Paris: OECD Publishing. <https://doi.org/10.1787/9789264266490-en>.
- Petscher, Y., Mitchell, A. M., & Foorman, B. R. (2015). Improving the reliability of student scores from speeded assessments: An illustration of conditional Item Response Theory using a computer-administered measure of vocabulary. *Reading and Writing: An Interdisciplinary Journal*, 28(1), 31–56. <https://doi.org/10.1007/s11145-014-9518-z>.
- Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement*, 74(3), 423–452. <https://doi.org/10.1177/0013164413504926>.
- Prediger, S., & Zindel, C. (2017). School academic language demands for understanding functional relationships: A design research project on the role of language in reading and learning. *EURASIA Journal of Mathematics Science and Technology Education*, 13(7b), 4157–4188. <https://doi.org/10.12973/eurasia.2017.00804a>.

- Rafferty, A., & Manning, C.D. (2008). Parsing three German treebanks: Lexicalized and unlexicalized baselines. In *ACL workshop on parsing German* (pp. 40–46). <https://doi.org/10.3115/1621401.1621407>.
- Rasinski, T. V. (2012). Why reading fluency should be hot! *The Reading Teacher*, 65(8), 516–522. <https://doi.org/10.1002/TRTR.01077>.
- Rose, N., Jonkmann, K., Hübner, N., Sälzer, C., Lüdtke, O., & Nagy, G. (2013). Durchführung und methodische Grundlagen der TRAIN-Studie [Implementation and methodological basics of the TRAIN study]. In K. Jonkmann, N. Rose, & U. Trautwein (Hrsg.), *Tradition und Innovation: Entwicklungsverläufe an Haupt- und Realschulen in Baden-Württemberg und Mittelschulen in Sachsen - Abschlussbericht für die Länder Baden-Württemberg und Sachsen* (pp. 77–102). Tübingen: Projektbericht an die Kultusministerien der Länder.
- Rose, N., von Davier, M., & Nagengast, B. (2016). Modeling omitted and not-reached items in IRT models. *Psychometrika*. <https://doi.org/10.1007/s11336-016-9544-7>.
- Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187–208). New York, NY: Springer.
- Schleppegrell, M. J. (2001). Linguistic features of the language of schooling. *Linguistics and Education*, 12(4), 431–459. [https://doi.org/10.1016/S0898-5898\(01\)00073-0](https://doi.org/10.1016/S0898-5898(01)00073-0).
- Schmidt, R. (1993). Awareness and second language acquisition. *Annual Review of Applied Linguistics*, 13, 206–226.
- Schrank, F. A., Mather, N., & McGrew, K. S. (2014). *Woodcock-Johnson IV tests of achievement*. Rolling Meadows, IL: Riverside.
- Shafteel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment*, 11, 105–126.
- Share, D. L. (2008). On the anglocentricities of current reading research and practice: The perils of overreliance on an “outlier” orthography. *Psychological Bulletin*, 134, 584–615.
- Temperley, D. (2007). Minimization of dependency length in written English. *Cognition*, 105, 300–333. <https://doi.org/10.1016/j.cognition.2006.09.011>.
- Tesnière, L. (1980). *Grundzüge der strukturalen Syntax [Fundamentals of structural syntax]* (herausgegeben und übersetzt von Ulrich Engel). Stuttgart: Klett-Cotta.
- Topping, K. J. (2006). Building reading fluency: Cognitive, behavioral, and socioemotional factors and the role of peer-mediated learning. In S. J. Samuels & A. E. Farstrup (Eds.), *What research has to say about fluency instruction* (pp. 106–129). Newark, NJ: International Reading Association.
- Turgay, K. (2010). *Der Zweitspracherwerb der deutschen Präpositionalphrase [Second language acquisition of the German prepositional phrase]*. Trier: Wissenschaftlicher Verlag.
- Turgay, K. (2011). Der Zweitspracherwerb des deutschen Kasus in der Präpositionalphrase [Second language acquisition of German case in the prepositional phrase]. *Zeitschrift für Germanistische Linguistik*. <https://doi.org/10.1515/zgl.2011.003>.
- Unsworth, S. (2016). Quantity and quality of language input in bilingual language development. In E. Nicoladis & S. Montanari (Eds.), *Lifespan perspectives on bilingualism* (pp. 136–196). Berlin: Mouton de Gruyter/APA.
- Ure, J. (1971). Lexical density and register differentiation. In G. Perren & J. L. M. Trim (Eds.), *Applications of linguistics* (pp. 443–452). London: Cambridge University Press.
- Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2010). *Test of silent reading efficiency and comprehension (TOSREC)*. Austin, TX: Pro-Ed.
- Wagner, W. (Ed.). (2013a). *Tradition und Innovation: Entwicklungsverläufe an Haupt- und Realschulen in Baden-Württemberg und Mittelschulen in Sachsen – Abschlussbericht für die Länder Baden-Württemberg und Sachsen [Tradition and innovation: Academic and psychosocial development in vocational track schools in the states of Baden-Württemberg and Sachsen]*. Tübingen: Projektbericht an die Kultusministerien der Länder.
- Wagner, W. (2013b). Durchführung und methodische Grundlagen der TRAIN-Studie [Implementation and methodological basics of the TRAIN study]. In K. Jonkmann, N. Rose, & U. Trautwein (Eds.), *Tradition und Innovation: Entwicklungsverläufe an Haupt- und Realschulen in Baden-Württemberg und Mittelschulen in Sachsen - Abschlussbericht für die Länder Baden-Württemberg und Sachsen* (pp. 77–102). Tübingen: Projektbericht an die Kultusministerien der Länder.
- Walczyk, J. J. (2000). The interplay between automatic and control processes in reading. *Reading Research Quarterly*, 35(4), 554–566.

- Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.