Special Issue

LANGUAGE TESTING

Language Testing
1–22


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0265532221991480
journals.sagepub.com/home/ltj
SAGE

Testing young foreign language learners' reading comprehension: Exploring the effects of working memory, grade level, and reading task

Tineke Brunfaut
Lancaster University, UK

Judit Kormos
University of Vienna, Austria & Lancaster University, UK

Marije Michel
University of Groningen, Netherlands

Michael Ratajczak
Lancaster University, UK

Abstract

Extensive research has demonstrated the impact of working memory (WM) on first language (L1) reading comprehension across age groups (Peng et al., 2018), and on foreign language (FL) reading comprehension of adults and older adolescents (Linck et al., 2014). Comparatively little is known about the effect of WM on young FL readers' comprehension, and even less within testing contexts. Young FL readers are still developing their L1 reading skills and general cognitive skills (e.g., attentional regulation abilities). Completing FL reading tests might be particularly taxing on their WM, and differences in WM capacity – as well as other learner and task characteristics – might create construct-irrelevant variance in test performance.

In this study we investigate the effects of WM, grade level, and reading task on young learners' FL reading test performances. Ninety-four young English language learners (Grades 6–7) in Hungary completed the TOEFL® Junior™ Comprehensive's reading test and a WM test battery. Our mixed-effects model predicted significantly higher comprehension accuracy among learners

Corresponding author:
Tineke Brunfaut, Department of Linguistics and English Language, Lancaster University, Lancaster, LA13YL, United Kingdom.
Email: t.brunfaut@lancaster.ac.uk

with higher WM capacity, and among learners in Grade 7 compared to learners in Grade 6. Reading task differences were not associated with significant comprehension accuracy differences. We discuss the implications of our findings for testing young learners' FL reading comprehension.

Keywords
FL reading, foreign language reading, grade level, L2 reading, language testing, reading, reading task, task, working memory, young learners

Literature review

The role of cognitive processes, text type and developmental factors in reading comprehension

Reading comprehension in another language is a complex cognitive process in which individuals' cognitive characteristics, skills, knowledge, and metacognitive processes interact with the text and the goals of the reading process (Enright et al., 2000; Khalifa & Weir, 2009). Recently, Francis et al. (2018) proposed a similar dynamically interactive view of reading in the field of L1 reading research. Importantly, Francis et al.'s Complete View of Reading from within an *interactive* lens (CVR*i*) takes into account developmental factors in order to capture how readers at various ages construct coherent mental representations of texts. Francis et al. argue that, in order to describe reading across ages, interactions between readers, texts, and the reading processes should be considered jointly. Current models of (testing) reading in a second or foreign language (L2),[1] however, have largely been informed by research on adult and academic L2 reading (Jeon & Yamashita, 2021), and therefore may not fully capture young learners' reading in another language. Thus, despite its L1 research base, Francis et al.'s CVR*i* seems worthwhile considering in the context of research on young foreign language learners' reading comprehension, in combination with what we know already about reading in another language.

The CVR*i* combines, in a unified model, previous theories on component skills of readers, text features that influence comprehension, and the development of reading comprehension through life stages (Francis et al., 2018). First, the component skills of reading in the CVR*i* model are based on Tunmer and Chapman's (2012) Modified Simple View of Reading. According to this view, the outcomes of reading comprehension are predicted by the joint interactive effect of general language comprehension skills and accurate and fluent written word decoding. In addition, vocabulary knowledge and the richness of lexical representations mediate the relationship between language comprehension and word decoding (Perfetti, 2007).

Second, in describing how different text features influence comprehension, the CVR*i* model draws on Kintsch's (1988) Construction-Integration model, which assumes that in reading comprehension lower-order, automatic, bottom-up and higher-order, top-down, conscious processes interact. Comprehension also involves textual processing at the surface, text, and situation model levels. At the surface level, readers process words and phrases contained in the text itself (Kintsch & Rawson, 2007). They rely on perceptual

processes and, following word recognition, assign words to their roles in phrases and sentences, a process known as parsing. At the text level, readers join the meaning of individual words to form propositions that represent the meaning of individual sentences. Propositions are then interconnected by the reader in a complex network, forming the microstructure of the text. Readers create the macrostructure by studying the coherence relations between propositions which they construct based on the microstructure. However, successful comprehension also requires the integration of the text-base representation into readers' background knowledge, which takes place at the situation model level (Kintsch, 1988). The situation model is influenced by factors such as readers' background knowledge, goals, personal experiences, and cognitive resources (Kintsch & Rawson, 2007).

Variation in text coherence and cohesion is also thought to influence comprehension (Kintsch, 1988). Texts that are not coherent may require the reader to use their background knowledge to establish coherence by using inferences (Van Dijk & Kintsch, 1983). Texts that differ in coherence and cohesive features make different processing demands on readers as they construct the situation model of the text (Kintsch & Rawson, 2007). Expository texts are assumed to be more taxing for readers because they tend to be informationally denser, use more complex syntactic structures, as well as less familiar academic or technical vocabulary than narratives (McNamara et al., 2012).

Third, the CVR*i* model also incorporates developmental perspectives. Findings in L1 research consistently show that as children's literacy skills develop, word level decoding skills become less accurate predictors of reading comprehension, and are replaced by vocabulary and background knowledge as important contributing factors to successful text comprehension (Oakhill & Cain, 2012). A recent meta-analysis by Peng et al. (2018) indicated that WM correlates more strongly with reading comprehension scores below Grade 4 than at/beyond Grade 4. This is because with the development of literacy skills, children become more efficient at word-level decoding and have higher levels of vocabulary knowledge and richer lexical representations at the higher grades. This makes lower-order reading skills more automatic and thereby less taxing on WM resources. Age-related differences in the quality of lexical representations also explain why children below Grade 4 understand narrative texts that tend to contain high-frequency words better than expository texts which often apply low-frequency and technical vocabulary (McNamara et al., 2012). Younger readers also find drawing inferences based on textual information and background knowledge more difficult than older students (Hannon & Daneman, 2009).

Readers and tasks in L2 language comprehension

As regards L2 reading processes, Cummins' (1979) influential *linguistic interdependence hypothesis* assumes that L1 and L2 literacy skills are closely interlinked and poor L1 skills are a critical contributor to L2 reading difficulties. In contrast, the *threshold hypothesis of linguistic competence* assumes that below a certain L2 proficiency level, L2 readers are not able to rely on their L1 reading skills to achieve successful L2 text comprehension (Alderson, 1984; Bernhardt & Kamil, 1995). A recent review by Pae (2019) of empirical studies on the relationship between L1 and L2 reading shows

substantial support for the linguistic interdependence hypothesis, whereas findings are contradictory regarding the existence of a linguistic threshold. An important point in the context of young L2 readers, however, is that their L1 literacy skills are still very much developing as well.

Furthermore, although L1 and L2 reading processes share several similarities, and L2 reading research has built considerably on L1 research, they also differ in crucial ways. One difference is that L2 readers often have a smaller vocabulary size, less rich lexical knowledge, and demonstrate slower speed in lexical access than L1 readers (Brysbaert et al., 2017; Geva & Farnia, 2012) which affects their comprehension levels. In fact, L2 vocabulary knowledge has consistently been found to be a strong predictor of L2 reading comprehension performance (e.g., Brunfaut, 2008; Van Gelderen et al., 2004). Furthermore, L2 speakers utilize their L1 to monitor their comprehension and accomplish metalinguistic functions, such as making observations about the text and reading behaviour, adjusting reading in response to text and reading demands (Upton & Lee-Thompson, 2001). L2 readers also use cognitive strategies, such as mental translation, to improve their L2 comprehension (Kern, 1994). In addition, Jeon and Yamashita's (2014) meta-analysis showed that variables such as L2 grammatical competence, socio-educational context, and task-related variables also influence L2 reading comprehension.

In learning and assessment contexts, L2 reading differences can stem from specific text and item characteristics as well as from the specific reading purposes set by the task. Task characteristics include variables such as the linguistic complexity, organisation, length, and genre of the reading input, or the item type used to elicit evidence of comprehension. With regard to the reading purposes, Khalifa and Weir (2009) proposed a framework which indicates that L2 readers can be engaged in different types of reading processes, such as careful versus expeditious reading and local versus global reading, depending on the goal of the reading task. The metacognitive processes of goal setting, monitoring, and remediating assist L2 readers in regulating their reading processes and achieving the required level of understanding depending on the task. From an empirical perspective, a meta-analysis by In'nami and Koizumi (2009) of test format effects, for instance, demonstrated that multiple-choice L2 reading tasks were easier than open-ended tasks given a number of conditions (e.g., stem equivalent items, high L2 proficiency). In another example, Brunfaut and McCray (2015) observed for adult English-L2 readers that, depending on where the gaps were created in gap-fill tasks, the tested construct constituted reading comprehension or vocabulary, and the cognitive and metacognitive reading processing differed.

Two further studies on the role of task in L2 reading are Löwenadler (2019) and Jung (2018). Löwenadler's research with young Swedish adults demonstrated that comprehension scores on short and long texts did not differ significantly and that a rational (semantic) deletion task was a better measure of L2-specific reading abilities than multiple-choice content questions. Jung's study revealed that L2 reading comprehension scores did not differ in simple task conditions when Korean university students had to answer multiple-choice items based on exploratory texts from the TOEFL iBT® and in complex task conditions when students additionally had to reorder jumbled passages of the texts. Little is known, however, about the role of tasks in young L2 learners' reading.

The role of working memory in reading

One of the most widely used WM models is Baddeley and Hitch's (1974) refined WM model (Repovš & Baddeley, 2006). Baddeley and Hitch originally proposed a three-component WM model comprising of a central executive (CE) aided by two storage-capacity-limited subsystems, a phonological loop, and a visuospatial sketchpad. The CE is assumed to be an attentional control system of limited processing capacity. The phonological loop stores and maintains verbal information, and the visuospatial sketchpad stores and maintains visual and spatial information. In 2000, Baddeley added another component – the episodic buffer – which is assumed to be a limited-capacity storage system that can integrate information from the phonological loop, the visuospatial sketchpad, and long-term memory. The role of the CE was also refined to include dividing attention between concurrent tasks, switching attention between different tasks, and inhibiting distracting material (Repovš & Baddeley, 2006).

WM is assumed to have an important function in reading comprehension because it assists in keeping processed bits of information active, updating readers' understanding with new information, and orchestrating all comprehension processes (van den Broek et al., 2016). A key cognitive component for efficient text processing are CE functions which are thought to help readers maintain focus while reading and inhibit irrelevant information (Oakhill et al., 2005). The potentially important role of differences in WM functioning is demonstrated in Peng et al.'s (2018) meta-analysis, showing a significant moderate correlation ($r = .29$) between L1 reading and WM.

WM resources are hypothesized to be involved differentially in processes which are automatic versus those that require conscious attention. For skilled readers, lower-order reading comprehension processes are automatized, and these processes therefore do not rely on WM. Consequently, readers have more WM resources available for maintaining information in active memory, integrating this information with relevant background knowledge and inhibiting redundant information. Thereby, they can create a more coherent situation model (Kendeou et al., 2014; Kintsch & Rawson, 2007). However, less skilled readers, and typically L1-speaking children below Grade 4, display lower levels of automaticity in word-level decoding. This can deplete their WM resources and result in difficulties in creating a text and situation model (Evans & Stanovich, 2013). The age-dependent role in WM also seems to be supported by Peng et al.'s (2018) meta-analysis, which revealed that WM plays a somewhat stronger role in L1 reading comprehension before Grade 4 ($r = .32$) than at or beyond Grade 4 ($r = .27$), also after controlling for a range of other variables ($\beta = .06$, $t = 2.57$, $p = .01$).

Based on Francis et al.'s (2018) CVR*i* model, text type and task difficulty can also be hypothesized to moderate the role of WM in text comprehension. Peng et al.'s (2018) meta-analysis examined this hypothesis, but found no significant difference in how strongly WM abilities predicted understanding narrative and expository texts, and the role of WM in understanding these two text types was also similar below and at/above Grade 4. They explained these findings by arguing that narrative texts used for assessing comprehension at/above Grade 4 often increase in difficulty because they require readers to draw more inferences, contain less frequent words, become longer, and use more complex sentence structure, which might mask any effects of WM.

With respect to L2 reading, a moderate effect size was reported in Linck et al.'s (2014) meta-analysis for the relationship between WM and L2 reading. Most research in this area, however, has focused on adult or older adolescent L2 readers. For example, Kormos and Sáfár's (2008) study explored the relationship between WM and L2 reading with somewhat older adolescents aged 15–16 in Hungary. While they did not detect a statistically significant relationship between L2 reading performance and phonological short-term memory, a moderately strong link was established between reading scores and complex WM capacity as measured by a backward digit span test.

As regards young L2 readers and their WM, some studies have been conducted in naturalistic, bilingual L2 learning contexts. Geva and Farnia (2012) found positive, significant correlations between WM, as measured by a backward digit span test, and reading comprehension scores of Grade 5 children in Canada ($r = .32$). In contrast, Raudszus et al.'s (2018) study with bilingual children in Grade 4 in the Netherlands detected no statistically significant direct links between backward digit span scores and Dutch reading comprehension ($r = .13$) or between inhibition measures and reading test scores ($r = .13$). Further statistical analyses in these studies revealed that when other predictors were added to the model of L2 reading comprehension, working memory either became a non-significant contributor (Geva & Farnia, 2012) or an indirect predictor of L2 reading outcomes (Raudszus et al., 2018; $\beta = .75$ via syntactic integration). In Raudszus et al.'s (2018) study, syntactic integration, assessed through a grammaticality judgement test, mediated the role of WM. However, when rate of growth in reading comprehension scores between Grade 4 and Grade 6 was examined by Farnia and Geva (2013), phonological short-term memory—measured by a non-word repetition task—was a significant correlate ($r = .30$) and an independent predictor of reading comprehension in a model with other oral language variables.

To the best of our knowledge, no previous studies have examined the role of WM in L2 reading outcomes of young learners in instructed foreign language learning contexts, the setting in which our study took place.

This study's aim

The above review indicates a gap in the L2 reading literature from a developmental perspective, given relatively scarce insights into young L2 learners' reading. In particular, the role of individual variables such as WM and of reading task variables on young L2 learners' comprehension processing is underexplored, especially in instructed L2 learning and assessment settings. To our knowledge, no previous research has examined the role of grade level, reading tasks, and WM abilities in one study. Therefore, this study set out to investigate the following research question:

> RQ. What is the role of grade level, reading task, and working memory capacity in the reading comprehension accuracy of young English as a foreign language (EFL) learners?

We thereby explore grade level as a developmental variable, as in many instructed, mainstream school contexts grade level is age-based and also associated with an expected

level of cognitive development. With increased grade level, children typically also receive more first language literacy and second language instruction. With regard to reading task, we define task as the combination of the input text and the item(s) associated with the text.

From a theoretical perspective, by examining the effects of this particular set of variables, our study is the first to test the applicability of the CVR*i* model in the field of (testing) L2 reading. Based on the above review of literature, our predictions are the following. With respect to grade level, based on cognitive and literacy developmental patterns in young learners, we hypothesized that young EFL learners' comprehension accuracy would be higher at higher grade levels in an instructed setting. With respect to reading task, we expected that task would affect reading accuracy, given findings demonstrating the role of task differences in L2 adult reading comprehension and given a set of tasks with a variety of characteristics. With respect to working memory capacity (WMC), based on earlier findings regarding the role of WMC in L2 adult and adolescent reading and in L1 and bilingual young learner reading, we hypothesized that WM would affect the English FL reading comprehension of young learners, albeit to a small extent.

This focus complements our earlier work where we looked into WM effects on L2 writing in young learners (Michel et al., 2019). Therefore, as part of a larger funded project on individual differences in young learner language assessment, we designed a study in which young EFL learners from two different levels of schooling completed a reading comprehension test, which contained four different tasks, and a WM test battery.

Methodology

Participants

Ninety-four young learners, aged 11–14 years (M_{age} = 12.22, SD = .78), participated in the study. Forty-five percent were boys and 55% were girls. Fifty-four percent were in Grade 6 (n = 51; M_{age} = 11.98, SD = .41) and 46% in Grade 7 (n = 43; M_{age} = 13.08, SD = .55). They were from two primary schools in Budapest (Hungary), where English is a foreign language. They had started learning English from Grade 1, with five English language lessons per week. From Grade 2, a content-based language instruction approach (CLIL) was added to this, with the children studying arts, music, science, and physical education through the medium of English in Grades 2–4, and history and science in Grades 5–8. The children's English proficiency ranged between CEFR A2–B2 (31% A2, 24% B1, and 45% B2), as determined by their results on the full TOEFL® Junior™ Comprehensive test-battery. We recruited participants from the population of English young learners in a CLIL setting in Hungary because they matched the target population of the TOEFL® Junior™ Comprehensive reading test, which constituted the reading measure in our study (see below). As described in ETS (2015) and So et al. (2015), the intended test-takers for this test are learners of English as a foreign language in non-English speaking countries (as, e.g., Hungary in our study), who need English for participation in an English-medium instructional environment (as, e.g., the CLIL setting in our study), who have more than a "basic" level of English-language ability (see, e.g., the CEFR levels of our participants reported above), and who are in the 11–15 years of age

range (as, e.g., Grades 6 and 7 in the Hungarian age-based school grade system; see age data of our participants above).

Instruments

Personal background questionnaire. Using the survey software Qualtrics, we administered a questionnaire to elicit information on the learners' gender, age, grade level, home language(s), residence abroad, length of learning English, and use of English outside the school context. We developed the questionnaire in English, and then translated and administered it in Hungarian.

Reading tasks. We assessed the learners' English reading comprehension ability by means of the reading test of the TOEFL® Junior™ Comprehensive test battery. This computer-based test aims to assess young learners' ability to understand texts for social, interpersonal, and navigational purposes, and also to understand academic texts from a range of genres and subjects (So et al., 2015). It contains four reading tasks which cover different language use domains and text genres, and 28 items which target a range of reading skills. See Table 1 for more information on the test version used in this study, and Table 2 for the linguistic complexity of each of the four input texts. These linguistic characteristics profiles were established by means of the ETS *TextEvaluator*® tool (https://www.ets.org/accelerate/ai-portfolio/textevaluator/).

Working memory tasks. To measure the participants' WMC, we identified WM tasks that (a) have been found suitable for use with young learners, given the developing nature of their cognitive functions (Gathercole et al., 2004) and (b) are as language independent as possible, given that our dependent variable (reading) is itself a language construct. In addition, we took into account the time the schools could make available for participation.

We administered three tasks, using the online tool Inquisit Web (www.millisecond.com), to measure aspects of the young learners' WMC. The tasks' instructions were presented to the learners in Hungarian. Two of the tasks, namely a visual forward and visual backward digit span task, were used to establish the young learners' storage and processing functions. In these tasks, the participant needs to recall a series of numbers presented on a computer screen, in order of display (forward task) or in reversed order (backward task), with the series increasing in length. Digit span tests seemed particularly suitable for our population, since these have specifically been identified as appropriate for testing the WMC of 11–14-year-olds (see Jarvis & Gathercole's (2003) review of WM tests for children and adolescents). A practical advantage of the visual span tasks was that we could administer them to groups of students at a time. In practice, we used digit span versions based on Woods et al. (2011; Experiment 1), with the scores on these tasks giving an estimate of the score a participant would obtain 50% of the time on the basis of overall performance during 14 trials.

The third task concerned the Symmetry Span task (Kane et al., 2004), which we selected to measure the task-switching functions of the young learners' WM. In this task, participants need to remember the location of a sequence of blocks (e.g., in a 4×4 grid).

Table 1. Reading test characteristics.

Reading input	Items	Skills targeted
Email	4 multiple-choice	• Identifying important supporting factual information • Discerning pronoun referent • Making inferences • Recognizing author's purpose or use of particular rhetorical structures
School news article	7 multiple-choice 1 insert sentence in text	• Comprehending main idea • Identifying important supporting factual information • Discerning meaning of words/expressions from context • Understanding figurative/idiomatic language from context • Understanding text coherence
Academic text	7 multiple-choice 1 insert sentence in text	• Comprehending main idea • Identifying important supporting factual information • Making inferences • Discerning meaning of words/expressions from context • Recognizing author's purpose or use of particular rhetorical structures • Understanding text coherence
Short story	7 multiple-choice 1 insert sentence in text	• Comprehending main idea • Identifying important supporting factual information • Making inferences • Discerning meaning of words/expressions from context • Recognizing author's purpose or use of particular rhetorical structures • Understanding text coherence

However, the task is interrupted by another task, which requires the participants to determine whether a black-and-white block pattern shown to them is symmetrical (Conway et al., 2005; see Figure 1). Scores on this task represent the sum of all items a participant accurately recalled in the correct order.

Procedures

Prior to data collection, all instruments were piloted with 14 young learners from the target population. This pilot indicated that the TOEFL® Junior™ Comprehensive test was suitable for the learners with respect to proficiency level, timing, and structure, and that the learners had the necessary computer and keyboard skills to complete this com-

Table 2. Reading input text characteristics.

Linguistic complexity	Measure	Email	School news	Academic text	Short story
Syntactic complexity	Syntactic complexity (+)	40	32	61	24
Lexical difficulty	Academic vocabulary (+)	39	52	58	40
	Word unfamiliarity (+)	46	61	55	43
	Concreteness (−)	58	47	51	66
Connections across ideas	Lexical cohesion (−)	31	32	55	24
	Interactive/Conversational style (−)	34	43	33	55
	Level of argumentation (+)	32	79	62	50
Organization	Degree of narrativity (−)	59	78	67	78
Overall text complexity	TextEvaluator complexity score (+)	670	740	790	459

(+) Higher values indicate higher complexity.
(−) Lower values indicate higher complexity.

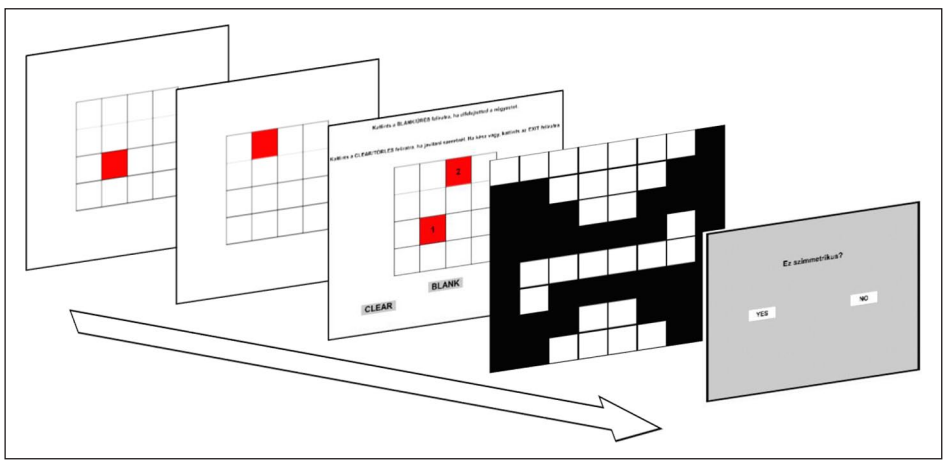

Figure 1. Symmetry span task.

puter-based test. The WM tasks also operated as intended, and the participants commented positively on the experience.

In the pilot as well as the main study, the young learners were first familiarized with the TOEFL® Junior™ Comprehensive test. The familiarization activities were conducted by the learners' regular English teachers during class time, by means of publicly available sample materials and the official test handbook.

The main study data were collected during two consecutive sessions. First, the learners completed the WM tasks: the two digit span tasks took approximately 5 minutes each and the Symmetry Span task approximately 10 minutes. Next, the young learners completed the TOEFL® Junior™ Comprehensive test, adhering to the test's standard regulations on timings and breaks. Finally, they completed the bio-data questionnaire. All instruments were administered in group in a computer room in the young learners' schools.

The study was approved by the relevant ethics review committee at the researchers' institution (Lancaster University). We obtained consent from both the young learners and their parents prior to data collection.

Analysis

To examine the factors that predicted the reading comprehension performance, we used Generalized Linear Mixed-Effects Models (GLMMs). We built these models using the glmer function in the lme4 package (Bates et al., 2015) in R (R Core Team, 2020). GLMMs were theoretically appropriate for this analysis, because we had item-level accuracy data that followed a binomial distribution (correct/incorrect). Thus, we had to model the likelihood of getting a comprehension question right, and GLMMs allowed us to do that.

The predictor variables in our models were as follows: School Grade (Grade 6 vs. Grade 7), Reading Task (an Email, School news, Academic, and Short story task), and WMC; for a study with similar variables, but exploring effects on writing, see Michel et al. (2019).

Results

Descriptive statistics

The descriptive statistics of learners' performances on the different reading tasks (Table 3) indicate that the Grade 6 and Grade 7 learners had the highest probability of getting an item right on the Email task. Conversely, the Grade 6 and Grade 7 learners had the lowest probability of getting a comprehension question right on the School news task. Table 3 also shows that the raw mean (*M*) differences between the tasks were relatively small for the young learners within each grade level. The raw mean differences for learners in different Grades, on the other hand, showed a pattern of higher probability of getting an item right across all tasks in Grade 7. The Cronbach alpha value for the reading test was .86.

The descriptive statistics for the WM tests (see Table 4) show that, on average, the learners had high digit span scores, despite their young ages. For reference, Jarvis and Gathercole (2003) reported mean scores of 6 (forward) and 5.5 (backward) for 14-year-olds, and Kormos and Sáfár (2008) 5.3 (backward) for 15–16-year-olds. The mean task switching score for our participants was 19. It should be noted, however, that our young learners differed considerably in their WMC, as demonstrated by the *SD*s. The two digit span tasks also seemed to partly tap into the same construct, given a relatively high

Table 3. Descriptive statistics: Reading tasks across Grade level.

Reading task	Grade	*M*	*SD*
Email	6	.76	.43
	7	.84	.37
	Total	.79	.41
School news	6	.69	.46
	7	.77	.42
	Total	.73	.45
Academic	6	.72	.45
	7	.83	.37
	Total	.77	.42
Short story	6	.73	.45
	7	.83	.38
	Total	.77	.42

Note: In each case, the minimum is 0 and the maximum is 1.

Table 4. Descriptive statistics: WM tests.

WMC test	*M*	*SD*
Forward Digit Span	6.08	.91
Backward Digit Span	5.58	.93
Symmetry Span	18.83	8.44

correlation between these two variables ($r = .60$, $p < .01$). Moderate correlations were also found between the Symmetry Span test and the Forward and Backward Digit Span tests ($r = .43$, $p < .001$; $r = .52$, $p < .001$). Therefore, we investigated the appropriateness of establishing a combined WM score. Principal component analysis confirmed that this was a suitable approach, and a composite score was established using regression factor scores. For a full description of this analysis, see Michel et al. (2019, p. 37).

Mixed-effects modelling results

As mentioned, to investigate the relationship between our predictor variables (Grade, Reading task, WMC) and TOEFL® Junior™ Comprehensive reading scores, we used GLMMs. To minimize the Type I error rate of our predictions, our models considered random variation between participants and test items (Jaeger, 2008). In other words, we deliberately added extra uncertainty into our models in order to account for between-participant and between-items variability. Adding random effects reduces the Type I error rate, as it lowers the probability of spuriously misattributing statistically significant effects to fixed effects of interest when they should actually be attributed to stochastic variation between participants or items (Yarkoni, 2019). Consequently, we added a

random intercept of participants to consider the differences in comprehension between participants, and a random intercept for items to take into account random variation between items within and between all the texts used in the study.[2]

Although we would have liked to test whether the effects of variation in WMC may have differed across grade levels and tasks, we were not able to do so reliably because our study was underpowered to do this in terms of *n*-size. In addition, while a model with the added interactions converged using the bound optimization by quadratic approximation algorithm (Powell, 2009), it did not converge using the glmer's default optimizer (Bates et al., 2015). Thus, we assumed that our study contained too few observations to reliably estimate the effects of theory-motivated interaction effects. Consequently, to avoid overfitting and to keep the model parsimonious, we retained the model without interaction terms. It thus remains an open question whether a model with interactions (exploring whether the effect of variation in WMC differs across grade levels and tasks) would approximate the reality better in this study than a model without interactions considered. The more complex interaction model still needs to be investigated, but a larger dataset is needed to do so.

Overall, we ran and evaluated a series of models to find a model that best fit our data (see Table 6 in the Appendix). We used the Likelihood Ratio Test (LRT; Baayen, 2008) to compare how well the simpler models fit our data in comparison with the more complex ones. The model comparisons in the Appendix show that accounting for extra uncertainty into our model, in order to account for between-participant and between-items variability, improved the model fit. In addition, the model that was found to approximate our data best considered additional uncertainty due to random differences in the slopes of the predicted effects of WMC. Specifically, this model took into account that the effect of WMC on reading comprehension accuracy can vary depending on the item being answered. The optimal model we arrived at (Model 6 in Table 6 in the Appendix), given our data, was

$$\text{Reading Comprehension Accuracy} \sim \text{Grade} + \text{Reading task} + \text{WMC} + \left(1|\text{Participants}\right) + \left(\text{WMC} + 1|\text{Items}\right)$$

where (*1*|*Participants*) denotes between-participant variability whereas (*WMC* + *1*|*Items*) denotes between-items variation and that the strength of the effect of WMC on reading comprehension accuracy can vary depending on the item being answered.

The optimal model accounted for 33.02% of the variance associated with reading comprehension accuracy (calculated using delta R^2 formula; Nakagawa et al., 2017). The random effects accounted for the majority of the variance (28.26%), indicating that a lot of variation in individuals' comprehension accuracy was owing to random differences between participants and test items. The rest of the variance in reading comprehension accuracy (4.76%) was accounted for by the predictor variables, indicating that some variation in reading comprehension accuracy was predicted by the effects of Grade, Reading task, and WMC.

Table 5 shows a summary of the optimal model and includes the log-odds estimates as well as Odds Ratio (OR) estimates and 95% profile confidence intervals (CIs) of OR estimates.

We found that on average Grade 7 learners were 2.212 [1.297, 3.851] times more likely to answer an item correctly than Grade 6 learners. In terms of predicted probabilities, our model estimated that Grade 7 learners had, on average, a 7.41% [2.947, 10.330][3] higher probability of answering items correctly compared to Grade 6 learners. This effect was statistically significant, indicating that Grade 7 Hungarian learners of English, in a CLIL setting, were more likely to answer the reading comprehension questions of the TOEFL® Junior™ Comprehension test correctly than their Grade 6 counterparts.

We also found that with one standard deviation (*SD*) increase in WMC from the mean, participants were 1.477 [1.103, 2.002] times more likely to answer an item correctly (see Figure 2). In terms of predicted probabilities, the probability of answering items correctly increased by, on average, 4.22% [1.175, 6.719] for one *SD* increase in WMC scores from mean WMC. Since the effect of WMC on comprehension was statistically significant, the effects of WMC on the probability of answering items of the TOEFL® Junior™ Comprehension reading test correctly, in the general CLIL population of Grade 6 and 7 Hungarian learners of English, is likely to be positive. In other words, Grade 6 and 7 learners with higher WMC were predicted to be more likely to answer reading comprehension questions correctly on the TOEFL® Junior™ Comprehension test compared to their counterparts with lower WMC.

In contrast to the effects of Grade and WMC, we found no significant differences between the four Reading tasks in the probability of correctly answering reading items. Consequently, the estimate of the effect of Reading task on the probability of answering comprehension questions correctly, in the general CLIL population of Grade 6 and 7 Hungarian learners of English, is not clear. This is because the estimates of plausible values, of the differences in comprehension between the different Reading tasks, are highly uncertain (see 95% CIs in Table 5).

Discussion

Overall, our participants achieved a relatively high level of performance on the TOEFL® Junior™ Comprehensive reading test (see descriptive statistics, Table 3), suggesting that this test was a suitable match for their proficiency level, but also that these young FL learners had a good level of understanding of English written texts considered appropriate for their age group. Although English is a foreign language in Hungary, and these learners were primarily acquiring the language in an instructed context, their schools had opted for a CLIL model wherein an increasing selection of other school subjects is taught through the medium of English. In practice, this means that, within their regular school context, these young FL learners are likely to be exposed to English texts on a daily basis, for a combination of reading-for-comprehension and reading-to-learn purposes, covering a wide range of topic areas through the various subjects. This solid amount and variety of reading input is likely to contribute to their FL reading development and explain their good levels of performance on the TOEFL® Junior™ Comprehensive reading test.

Table 5. Summary of the final model.

Fixed effects	Estimate	OR	95% OR CIs [2.5%, 97.5%]	Standard error	*z*-value	*p*
(Intercept)	1.769	5.863	[2.000, 17.266]	.523	3.380	.001
Grade: 7	.794	2.212	[1.297, 3.851]	.271	2.927	.003
Reading task: School news	−.703	.495	[.147, 1.634]	.589	−1.194	.233
Reading task: Academic passage	−.412	.663	[.186, 2.157]	.618	−.665	.506
Reading task: Short story	−.420	.657	[.195, 2.157]	.589	−.713	.476
WMC	.390	1.477	[1.103, 2.002]	.148	2.631	.009
Random effects (Intercepts)	Random Slopes	Variance	*SD*	Correlations		
Participants		1.352	1.163			
Items		.804	.897			
	WMC	.082	.286	.170		

Note: Grade: 6 is the reference level for Grade; Task: Email is the reference level for Task; Working memory refers to mean centred and standardised working memory capacity scores. OR refers to odds ratio; CIs refers to profile confidence intervals.

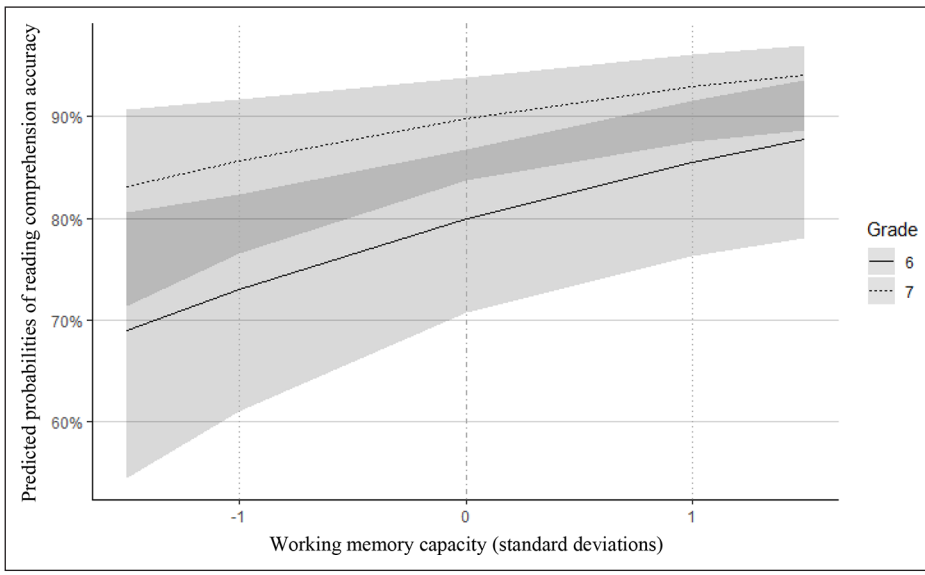

Figure 2. The effects of variation in working memory capacity on reading comprehension accuracy.

A specific goal of our study was to explore factors which specifically affect young learners' FL reading, by which we aimed to help address current gaps in developmental perspectives within L2 research (see literature review). Based on Francis et al.'s (2018)

CVR*i* and our review of factors that play a role in older adolescent and adult L2 reading, we set out to examine the role of Grade level, reading tasks, and WM abilities in young FL learners' reading comprehension performances. Using a GLMM model, we found that these variables—Grade, reading tasks, and WMC—accounted for 4.76% of the variance in the young learners' FL reading comprehension accuracy. The limited amount of variance explained by the variables under focus seems to lend support to findings from previous studies which have shown that other factors, such as vocabulary and syntactic knowledge, are key predictors of L2 reading performance (Jeon & Yamashita, 2014).

With respect to the variables explored, we did not find any significant differences in reading comprehension performance *between the reading tasks*. At first sight, this was unexpected, given earlier findings on the role of tasks in the testing of adult L2 reading (see literature review) and given the variety in input texts in the TOEFL® Junior™ Comprehensive reading test. Namely, the four texts differed in genre, topic, and linguistic complexity (see Table 2), and were designed to target different overall reading purposes (social, interpersonal, navigational, and academic reading purposes). At the same time, however, the four reading tasks were all selected-response formats, with all items being multiple-choice, apart from one item each in the School news, Academic, and Short story tasks (this item required the test-takers to insert a missing sentence in the relevant space in the input text). Furthermore, the set of items in each reading task targeted a great mix of reading subskills, and there was overlap in subskills targeted between the four reading tasks (see Table 1). Thus, while the four reading tasks were distinct in some respects (especially in terms of input texts), they also shared a number of features (especially as related to the items and what these targeted). Therefore, since reading scores are the result of the interaction between texts and items (Alderson, 2000), the text difficulty differences between the four reading tasks might have been balanced out by shared item features.

The levels of reading comprehension performance *between learners from the two Grade levels* differed significantly. Namely, Grade 7 learners had a higher comprehension accuracy than Grade 6 learners. For the present study's population and context, it can be argued that Grade level is a proxy for age-related cognitive development and literacy skills development, as well as for the young learners' relative amount of exposure to English. That is, first of all, entry and progression in the Hungarian schooling system are mostly age based, with each grade level constituting a cohort of learners of a similar age (and age-associated cognitive development), and different in age (and age-associated cognitive development) from learners of another grade. Second, with an additional year of schooling, learners in a higher grade will have received an extra year of literacy training, and are thus likely to have higher literacy skills. Third, the Hungarian context constitutes an English foreign language setting, with comparatively limited out-of-school exposure and production opportunities in daily life. The amount of exposure to English is therefore largely governed by the instructed setting, with Grade 7 learners having had one more year of English exposure and acquisition opportunities in the CLIL setting than Grade 6 learners. Our Grade-related findings thus suggest that the TOEFL® Junior™ Comprehensive reading test is successful in detecting developmental differences for reading tasks which target understanding of a range of texts of different genres and linguistic complexities, reading purposes, and reading skills. It also suggests that reading comprehension of a variety of texts can improve with one additional year of instruction.

The *effect of WMC* on reading comprehension performance was significant, and the effects of WMC on the probability of answering items of the TOEFL® Junior™ Comprehension reading test correctly, in the general CLIL population of Grade 6 and 7 Hungarian learners of English, is likely to be positive. In other words, Grade 6 and 7 learners with higher WMC were predicted to be more likely (by an average of 4.22% per *SD* increase in WMC scores) to answer reading comprehension questions correctly on the TOEFL® Junior™ Comprehension test compared to their counterparts with lower WMC. This confirmed our hypothesis that working memory plays a significant, albeit small, role in young EFL learners' reading comprehension in instructed settings. The results suggest that from a developmental perspective, WM plays a relatively minor role when decoding processes are more automated, as discussed in the L1 reading literature (Oakhill & Cain, 2012; Peng et al., 2018). Since the group of young FL learners in this study had already been learning English for 6 to 7 years, several years of which in a CLIL environment with considerable amounts of exposure to English texts, they might have had a good level of efficiency in lower-level reading processes (including automated decoding) in English. However, the assumed functions of WM—keeping processed bits of information active, updating readers' understanding with new information, orchestrating all comprehension processes, helping to maintain focus, and inhibiting irrelevant information (Oakhill et al., 2005; van den Broek et al., 2016)—may thus have some role to play in the higher-order reading processes of this population.

Implications and limitations

With respect to reading theory, by examining the effects of Grade, reading task, and WMC, our study was the first to test the applicability of Francis et al.'s (2018) Complete View of Reading (CVR*i*) in the field of (testing) L2 reading comprehension. Our findings indicate that WMC and developmental stage (as operationalised by Grade) are likely to have an effect on L2 reading comprehension, thus providing support for the CVR*i* model. Although we found that these variables explained some variance in young learners' FL reading comprehension, the low percentage of variance accounted for seems to confirm the importance of other, previously identified factors such as FL vocabulary knowledge in componential L2 reading models.

With respect to assessment, our findings indicate that there is a small advantage for young FL learners with higher WMC on the TOEFL® Junior™ Comprehensive reading test. We should emphasize, however, that we did not control for a variable such as L2 vocabulary knowledge in our study, which has repeatedly been shown to be a strong predictor of L2 reading comprehension. As working memory and L2 vocabulary have been shown to correlate strongly in adolescent L2 learners (e.g., Lockiewicz & Jaskulska, 2015), it is unclear whether the effect of WMC in our study would prevail if L2 vocabulary knowledge was taken into account. In addition, although we cannot exclude a potential small testing method advantage for the high WMC learners, it is possible that our study's result is due to the underlying higher level of reading competence these high WMC learners have attained.

The TOEFL® Junior™ Comprehensive reading test seems to display sufficient sensitivity to detect developmental differences across educational grade level, and therefore might be used to assess students' progress in CLIL contexts similar to the investigated Hungarian one. The findings also indicate that students' reading comprehension across these four types

of input texts when assessed with similar types of items is relatively uniform. This result might show that owing to the CLIL context, students are familiar with the assessed genres of texts and on average can understand them with a high level of accuracy. This can serve as evidence for the purposeful selection of the target language domain of the CLIL context in TOEFL® Junior™ Comprehensive reading tasks. In sum, the TOEFL® Junior™ Comprehensive reading test was found to be a suitable reading comprehension measure for our population of Hungarian-L1, English-FL young learners with overall proficiency levels ranging between CEFR A2–B2. It also confirms the importance of careful task design for testing the reading comprehension skills of FL young learners; both the texts young learners are asked to read and the comprehension questions that are being posed on these texts should align with the young learners' language use domain and setting.

While our research contributes importantly to theory and practice in testing L2 reading, we want to acknowledge two important limitations. Although our study was conducted in a foreign language context (in fact one with languages from distinct language families: Hungarian L1, English FL), the young learners were learning English in a context with a CLIL pedagogy. The findings may thus not generalize to other FL contexts with more restricted learning of English as a subject only. In addition, given limitations to the size of our dataset, we were unable to explore interaction effects between the variables, and thus we only tested the model wherein each of the effects of Grade, Reading task and WMC were constant (i.e., we assumed that the effects of WMC did not vary depending on the Reading task or participants' Grade level, and vice versa, because our study did not have enough power to allow us to assume otherwise). The possibility of interaction effects was left untested, and warrants further research with a larger sample of participants. Such results would be important for understanding our target population, as well as other populations, and would help researchers and practioners better understand reading comprehension tasks (the texts and the items) used with young learners.

Acknowledgements

We would like to thank all the young learners who took part in the study, as well as their parents, teachers and schools who supported the project with great enthusiasm. We owe our thanks to Dr Gabriella Dóczi Vámos, Stella Varga, and Orsolya Szatzker who acted as local research assistants in Hungary. We would also like to express our gratitude to the research committee for the TOEFL® Young Students Series Research Grants (2016) and the staff at ETS, in particular, Veronika Timpe-Laughlin, for their support and prompt answers to our queries.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by Educational Testing Service (ETS), USA, under a Committee of Examiners and the Test of English as a Foreign Language Young Students research grant. ETS does not discount or endorse the methodology, results, implications or opinions presented by the researcher(s).

ORCID iDs

Tineke Brunfaut https://orcid.org/0000-0001-8018-8004
Judit Kormos https://orcid.org/0000-0002-2643-7222
Marije Michel https://orcid.org/0000-0003-1426-4771
Michael Ratajczak https://orcid.org/0000-0003-0562-5328

Notes

1. We use "L2" as encompassing second and foreign language, aligning with its use in SLA, covering both of these in target-language/immersion settings, given blurred distinctions.
2. Although items were, by design, nested within the tasks, this structure was too complex to be supported by the data as nesting items within tasks led to non-convergence. Consequently, we modelled a random intercept of items only. In other words, we assumed that each item within and between all texts could vary in difficulty. In addition, for the same reason of non-convergence, we did not nest participants within schools.
3. Changes in predicted probabilities were calculated using the following formula: (exp(sum of log odds of an effect and intercept) / (1 + exp(sum of log odds of an effect and intercept))) − (exp(log odds of the intercept) / (1 + exp(log odds of the intercept))). For example, the mean difference in predicted probabilities, between Grade 7 versus Grade 6 learners, was calculated as follows: (exp(2.5623) / (1 + exp(2.5623))) − (exp(1.7686) / (1 + exp(1.7686))) = 0.07411203 = 7.41%.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press.

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge University Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. A. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 47–89). Academic Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bernhardt, E. B., & Kamil, M. L. (1995). Interpreting relationships between L1 and L2 reading: Consolidating the linguistic threshold and the linguistic interdepence hypotheses. *Applied Linguistics*, *16*(1), 15–34. https://doi.org/10.1093/applin/16.1.15

Brunfaut, T. (2008). *Foreign language reading for academic purposes* [Unpublished doctoral dissertation]. University of Antwerp, Belgium.

Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processing whilst completing reading tasks: A mixed-methods eye-tracking and stimulated recall study*. ARAGs Research Reports Online, AR/2015/01. British Council. https://www.britishcouncil.org/sites/default/files/brunfaut_and_mccray_report_final_0.pdf

Brysbaert, M., Lagrou, E., & Stevens, M. (2017). Visual word recognition in a second language: A test of the lexical entrenchment hypothesis with lexical decision times. *Bilingualism: Language and Cognition*, *20*(3), 530–548. https://doi.org/10.1017/S1366728916000353

Conway, A. R. A., Kane, M. J., Bunting, M. F. D., Zach Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin and Review*, *12*, 769–786. https://doi.org/10.3758/BF03196772

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, *49*(2), 222–251. https://doi.org/10.3102/00346543049002222

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedle, M. (2000). *TOEFL 2000 reading framework: A working paper* (TOEFL Monograph No. MS-17). ETS. https://www.ets.org/research/policy_research_reports/publications/report/2000/iciv

ETS (2015). *Handbook for the TOEFL® Junior™ Comprehensive Test*. Educational Testing Service.

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Farnia, F., & Geva, E. (2013). Growth and predictors of change in English language learners' reading comprehension. *Journal of Research in Reading*, *36*(4), 389–421. https://doi.org/10.1111/jrir.12003

Francis, D. J., Kulesz, P. A., & Benoit, J. S. (2018). Extending the Simple View of Reading to account for variation within readers and across texts: The Complete View of Reading (CVR*i*). *Remedial and Special Education*, *39*(5), 274–288. https://doi.org/10.1177/0741932518772904

Gathercole, S. E., Pickering, S .J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, *40*(2), 177–192. https://doi.org/10.1037/0012-1649.40.2.177

Geva, E., & Farnia, F. (2012). Developmental changes in the nature of language proficiency and reading fluency paint a more complex view of reading comprehension in ELL and EL1. *Reading and Writing*, *25*(8), 1819–1845. https://doi.org/10.1007/s11145-011-9333-8

Hannon, B., & Daneman, M. (2009). Age-related changes in reading comprehension: an individual-differences perspective. *Experimental Aging Research*, *35*(4), 432–456. https://doi.10.1080/03610730903175808

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*(2), 219–244. https://doi.org/10.1177/0265532208101006

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Jarvis, H. L., & Gathercole, S. E. (2003). Verbal and non-verbal working memory and achievements on national curriculum tests at 11 and 14 years of age. *Educational and Child Psychology*, *20*(3), 123–140.

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, *64*(1), 160–212. https://doi.org/10.1111/lang.12034

Jeon, E. H., & Yamashita, J. (2021). Measuring L2 reading. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 265–274). Routledge.

Jung, J. (2018). Effects of task complexity and working memory capacity on L2 reading comprehension. *System*, *74*, 21–37. https://doi.org/10.1016/j.system.2018.02.005

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology General*, *133*(2), 189–217. https://doi.org/10.1037/0096-3445.133.2.189

Kendeou, P., Van Den Broek, P., Helder, A., & Karlsson, J. (2014). A cognitive view of reading comprehension: Implications for reading difficulties. *Learning Disabilities Research and Practice*, *29*(1), 10–16. https://doi.org/10.1111/ldrp.12025

Kern, R. G. (1994). The role of mental translation in second language reading. *Studies in Second Language Acquisition*, *16*(4), 441-–61. https://doi.org/10.1017/S0272263100013450

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge University Press.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review*, *95*, 163–182. https://doi.org/10.1037/0033-295X.95.2.163

Kintsch, W., & Rawson, K. A. (2007). Comprehension. In M. J. Snowling & C. Hulme (Eds.). *The science of reading: A handbook* (pp. 209–226). Blackwell.

Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition*, *11*(2), 261–271. https://doi.org/10.1017/S1366728908003416

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin and Review*, *21*, 861–883. https://doi.org/10.3758/s13423-013-0565-2

Lockiewicz, M., & Jaskulska, M. (2015). Mental lexicon, working memory and L2 (English) vocabulary in Polish students with and without dyslexia. *CEPS Journal*, *5*(1), 71–89. https://files.eric.ed.gov/fulltext/EJ1128848.pdf

Löwenadler, J. (2019). Patterns of variation in the interplay of language ability and general reading comprehension ability in L2 reading. *Language Testing*, *36*(3), 369–390. https://doi.org/10.1177/0265532219826379

McNamara, D. S., Graesser, A. C., & Louwerse, M. M. (2012). Sources of text difficulty: Across genres and grades. In J. P. Sabatini, E. R. Albro & T. O'Reilly (Eds.), *Measuring up: Advances in how to assess reading ability* (pp. 89–116). Rowman and Littlefield Education.

Michel, M., Kormos, J., Brunfaut, T., & Ratajczak, M. (2019). The role of working memory in young second language learner's written performances. *Journal of Second Language Writing*, *45*, 31–45. https://doi.org/10.1016/j.jslw.2019.03.002

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society, Interface*, *14*, 1–11. https://doi.org/10.1098/rsif.2017.0213

Oakhill, J. V., & Cain, K. (2012). The precursors of reading ability in young readers: Evidence from a four-year longitudinal study. *Scientific Studies of Reading*, *16*(2), 91–121. https://doi.org/10.1080/10888438.2010.529219

Oakhill, J., Hartt, J., & Samols, D. (2005). Levels of comprehension monitoring and working memory in good and poor comprehenders. *Reading and Writing*, *18*(7–9), 657–686. https://doi.org/10.1007/s11145-005-3355-z

Pae, T.-I. (2019). A simultaneous analysis of relations between L1 and L2 skills in reading and writing. *Reading Research Quarterly*, *54*(1), 109–124. http://dx.doi.org/10.1037/bul0000124

Peng, P., Barnes, M., Wang, C., Wang, W., Li, S., Swanson, H. L., Dardick, W., & Tao, S. (2018). A meta-analysis on the relation between reading and working memory. *Psychological Bulletin*, *144*(1), 48–76. https://doi.org/10.1037/bul0000124

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357–383. https://doi.org/10.1080/10888430701530730

Powell, M. J. D. (2009). *The BOBYQA algorithm for bound constrained optimization without derivatives*. Department of Applied Mathematics and Theoretical Physics, Cambridge University, DAMTP 2009/NA06. http://www.damtp.cam.ac.uk/user/na/NA_papers/NA2009_06.pdf

R Core Team (2020). *R: A language and environment for statistical computing*. The R Foundation. http://www.R-project.org/

Raudszus, H., Segers, E., & Verhoeven, L. (2018). Lexical quality and executive control predict children's first and second language reading comprehension. *Reading and Writing*, *31*(2), 405–424. https://doi.org/10.1007/s11145-017-9791-8

Repovš, G., & Baddeley, A. (2006). The multi-component model of working memory: Explorations in experimental cognitive psychology. *Neuroscience*, *139*(1), 5–21. https://doi.org/10.1016/j.neuroscience.2005.12.061

So, Y., Wolf, M. K., Hauck, M. C., Mollaun, P., Rybinski, P., Tumposky, D., & Wang, L. (2015). TOEFL Junior® Design Framework. *ETS Research Report Series*, *2015*(1), 1–45. https://doi.org/10.1002/ets2.12058

Tunmer, W. E., & Chapman, J. W. (2012). The simple view of reading redux: Vocabulary knowledge and the independent components hypothesis. *Journal of Learning Disabilities*, *45*(5), 453–466. https://doi.org/10.1177/0022219411432685

Upton, T. A., & Lee-Thompson, L. C. (2001). The role of the first language in second language reading. *Studies in Second Language Acquisition*, *23*(4), 469–495. https://doi.org/10.1017/S0272263101004028

van den Broek, P., Mouw, J. M., & Kraal, A. (2016). Individual differences in reading comprehension. In P. Afflerbach (Ed.), *Handbook of individual differences in reading: Reader, text, and context* (pp. 138–150). Routledge.

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. Academic Press.

Van Gelderen, A., Schoonen, R., De Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2004). Linguistic knowledge, processing speed, and metacognitive knowledge in first- and second-language comprehension: A componential analysis. *Journal of Educational Psychology*, *96*(1), 19–30. https://doi.org/10.1037/0022-0663.96.1.19

Woods, D. L., Kishiyama, M. M., Yund, E. W., Herron, T. J., Edwards, B., Poliva, O., Hink, R. F., & Reed, B. (2011). Improving digit span assessment of short-term verbal memory. *Journal of Clinical and Experimental Neuropsychology*, *33*, 101–111. https://doi.org/10.1080/13803395.2010.493149

Yarkoni, T. (2019). *The generalizability crisis*. PsyArXiv Preprints. https://doi.org/10.31234/osf.io/jqw35

Appendix

Table 6. Model comparisons.

Model	Model specification		Notes	Converged	Deviance	*p*
	Predictors	Random effects				
Selection of predictor variables						
1	Intercept only	(1|Participants) + (1|Items)	–	Yes	2434.9	–
2	Grade + Task + WMC	(1|Participants) + (1|Items)	–	Yes	2417.1	.003
3	Grade × Task × WMC	(1|Participants) + (1|Items)	Interactions model	No	–	–
Evaluation of the utility of random intercepts (against Model 2)						
4	Grade + Task + WMC	(1|Items)	–	Yes	2668.2	–
5	Grade + Task + WMC	(1|Participants)	–	Yes	2608.6	–
2	Grade + Task + WMC	(1|Participants) + (1|Items)	–	Yes	2417.1	< .001
Evaluation of the utility of random slopes (against Model 2)						
2	Grade + Task + WMC	(1|Participants) + (1|Items)	–	Yes	2434.9	–
6	Grade + Task + WMC	(1|Participants) + (WMC + 1|Items)	Optimal model	Yes	2408.4	.013
7	Grade + Task + WMC	(1|Participants) + (Grade + 1|Items)	Singular autocorrelations	Yes*	–	–
8	Grade + Task + WMC	(Task + 1|Participants)– + (1|Items)		No	–	–

Notes: Task refers to Reading task. Model 6 was the optimal model that was used in the primary analysis. *p* was calculated using the likelihood ratio test whereby the more complicated models were evaluated against the simpler models. Significant *p* is indicating an improvement in model fit of the more complicated model versus the simpler model. The smaller the deviance the better the model approximates reality. *Model 7 did converge, but its estimates cannot be trusted as singularity means that the variance of some of the components of the model will approach infinity.