BRITISH COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

TIME-EXTENSION AND THE SECOND LANGUAGE READING PERFORMANCE OF CHILDREN WITH DIFFERENT FIRST LANGUAGE LITERACY PROFILES

AR-G/2019/1

Judit Kormos and Michael Ratajczak
Lancaster University

ARAGs RESEARCH REPORTS ONLINE
SERIES EDITOR: VIVIEN BERRY
ISSN 2057-5203

ABSTRACT

Currently no study has systematically investigated how second language learners (L2) with specific learning difficulties (SpLDs) benefit from extended time in L2 assessment. Research in this area is needed because judgements about time extensions are often based on intuitions rather than on research evidence. This study investigated the effect of different timing conditions on the L2 reading performance of adolescent learners of English who demonstrate different first language (L1) literacy profiles. It aimed to uncover whether Hungarian L2 learners who have below average L1 reading comprehension and word-decoding skills, which can be indicative of SpLDs, gain from time extension differentially in the reading component of the *Aptis for Teens* test from those whose L1 skills are in the average or above average range.

Our generalised linear mixed-effects model predicted no significant effect of the time extension and no interaction between time extension, L1 skills and test tasks. This suggests that time extension did not boost students' scores and did not confer a differential advantage for students with low-level L1 skills either. Our results confirm the importance of universal test design and providing a generous margin of around 50% extra time from the mean test population completion time for all test-takers.

Authors

Judit Kormos

Judit Kormos is a Professor in Second Language Acquisition at Lancaster University. She was a key partner in the award-winning DysTEFL project sponsored by the European Commission and is a lead educator in the Dyslexia and Foreign Language Teaching massive open online learning course offered by FutureLearn. She is co-author of the book, *Teaching Languages to Students with Specific Learning Differences*, with Anne Margaret Smith. Judit has published widely on the effect of dyslexia on learning additional languages, including a book entitled, *The Second Language Learning Processes of Students with Specific Learning Difficulties*. She is the author of multiple research papers that investigate the role of cognitive factors in second language acquisition.

Michael Ratajczak

Michael Ratajczak is a specialist in multilevel modelling. He completed an MSc in Psychological Research Methods and a BSc in Psychology in Education at Lancaster University. Currently, Michael is a PhD student in the Department of Linguistics and English Language at Lancaster University, funded by the UK's Economic and Social Research Council (ESRC) and by the UK's National Health Service (NHS).

CONTENTS

List of tables

List of figures

1. INTRODUCTION

The use of international proficiency tests has become increasingly widespread because evidence of second language (L2) competence is frequently required in academic and professional contexts, as well as for immigration purposes. Therefore, the population of test-takers has increased substantially and so has the impact of tests on candidates' life chances, education and career prospects and immigration status. This change in test impact has highlighted the crucial role of test fairness. A fair and valid test ensures that each and every test-taker has equal chances to demonstrate their abilities in the assessment process. Although the ultimate aim is the design of tests that are universally accessible to all candidates including those with disabilities, in practice, test-takers with disabilities often need special arrangements that assist them to perform to the best of their knowledge in a test.

A large group of test takers with disabilities are those with specific learning difficulties (SpLDs), who constitute around 10–15% of the global population (OECD, 2012). The number of individuals with SpLDs taking language proficiency tests has grown substantially (Tsagari & Spanoudis, 2013). One of the most commonly used special arrangements offered to candidates with SpLDs is extended time which allows students a longer period to complete the assessment tasks. Extended time is hypothesised to assist test-takers with SpLDs who are often slow readers in their first language (L1) and who tend to be characterised by slower speed of information processing (Haladyna & Downing, 2004).

Extended time is often granted to students with SpLDs in L1 literacy assessment contexts. The benefits of time extension in L1 have been extensively studied, but previously no study has systematically investigated whether L2 learners gain from such an arrangement. Research in this area is needed because judgements about awarding time extension to students with SpLDs are often made based on intuition rather than on research evidence. Furthermore, the appropriate length of time for the completion of a test is also related to validity issues. If extended time results in score gains regardless of the SpLD status of the candidates, then test-takers in general may not be performing to the best of their knowledge under standard administration procedures. This can present a threat to the validity of the test as it influences the accuracy of conclusions one can draw about students' abilities based on the test results.

2. THEORETICAL BACKGROUND

2.1 The relationship between first language skills and second language reading

Research on reading across a variety of languages and orthographies demonstrates that using sounds of a language to process spoken and written texts plays a key role in reading in any language (cf. Perfetti, Zhang, & Berent, 1992). Therefore, phonological decoding is a universal predictor of the rate and ultimate attainment in reading development. In their *common underlying processes framework,* Geva and Ryan (1993) also argue that reading development in both monolingual and bilingual children is influenced by a key set of individual difference variables. A similar position is advocated by the *linguistic interdependence hypothesis* (Cummins, 1979) which posits that L1 and L2 literacy skills are strongly inter-related. According to this hypothesis, L1 reading skills and strategies are automatically transferred to L2 reading, and the major cause of L2 reading difficulties is poor L1 skills. 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 (e.g. Alderson, 1984; Bernhardt & Kamil, 1995).

Currently there is no conclusive evidence as regards the existence of a linguistic threshold for the successful transfer of L1 skills (for a recent review see Pae, 2019). However, it is important to acknowledge that in addition to L1 skills, other factors such as L2 vocabulary and grammar knowledge, socio-educational context, and task-related variables also influence L2 reading performance (see Jeon & Yamashita's (2014) meta-analysis).

Among one of the key predictors of reading performance in childhood is phonological awareness, that is, the ability to recognise, identify and manipulate phonological units of various size such as syllables, onset, rhymes and phonemes (Ziegler & Goswami, 2005). Phonological awareness, however, develops as children learn to read and it loses its importance as a contributor to text comprehension (Landerl et al., 2013). Its role is taken over by word naming speed (for a review see Kirby, Georgiou, Martinussen & Parrila, 2010), which is a measure of an individual's ability to access appropriate lexical representations under time constraints. Phonological awareness and rapid automated naming in L1 have also been found to be relatively good predictors of L2 reading skills of bilingual children (Erdos, Genesee, Savage & Haigh, 2014).

In the beginning stages of literacy development, efficient word-level decoding is essential for children to be able to understand sentences and longer texts, and both mono- and bilingual children demonstrate substantial variation in this low-level reading skill (Geva, 2000; Gough & Tunmer, 1986; Tunmer & Chapman, 2012). Tests of word-level decoding, such as timed and non-timed word reading, administered in children's L1 have been found to be relatively good predictors of reading attainment in L2 (Alderson, Haapakangas, Huhta, Nieminen & Ullakonoja 2015; Kormos, Kosak-Babuder & Pizorn, 2019; Van Gelderen, Schoonen, De Glopper, Hulstijn, Simis, Snellings & Stevenson, 2004). Phonological processing difficulties and slower speed of word naming are key characteristics of dyslexic readers, who exhibit word-level decoding problems and who will be discussed in more detail in the next section.

2.2 Specific learning difficulties

One of the most up-to-date definitions of learning difficulties is formulated in the fifth edition of the *Diagnostic and Statistical Manual* (DSM-5) of the American Psychiatric Association (APA, 2013). In DSM-5, *specific learning disorder*s, which is a term equivalent to SpLDs in the UK, are subdivided into further categories, such as "specific learning disorder in reading" and "specific learning disorder in written expression". Within the category of specific learning disorder in reading, word-level decoding problems (dyslexia) and higher-level text comprehension problems (specific reading comprehension impairment) are distinguished. Reading-related SpLDs tend to be caused by underlying weaknesses in the areas of phonological processing, working memory, attention regulation and processing speed (for a comprehensive discussion see Kormos, 2017).

As discussed above, the basic cognitive factors that account for L1 and L2 language and literacy development overlap, and L1 skills serve as an important foundation for L2 development (for a review see Kormos, 2017). Similar to the common underlying processes framework (Geva & Ryan, 1993) and the linguistic interdependence hypothesis (Cummins, 1979), Sparks and Ganschow's (1993) *Linguistic Coding Differences Hypothesis* assumes that the fundamental cognitive reasons for low achievement in L2 are similar to those factors that can explain L1 literacy problems. Nevertheless, there is conflicting evidence whether struggling L2 learners also face challenges in their L1 literacy, and whether L1 reading problems are associated with L2 learning challenges.

A number of research findings indicate that dyslexic-type difficulties tend to be associated with L2 reading comprehension problems. Both Norwegian (Helland & Kaasa, 2005) and Hungarian children with an official diagnosis of dyslexia (Kormos & Mikó, 2010) were found to achieve lower scores on L2 English word reading than non-dyslexic children. Hungarian L2 learners with SpLDs also obtained lower scores on a sentence comprehension test than their peers matched for age and the number of years of English language instruction (Kormos & Mikó, 2010).

Geva, Wade-Woolley and Shany (1993) in Canada, Crombie (1997) in Scotland and Sparks and Ganschow (2001) in the United States also established that L2 learners with dyslexic-type reading difficulties experienced challenges in L2 reading. In a recent study, Kormos et al. (2018) found that low-level L1 skills, including phonological awareness, timed word and non-word reading and orthographic skills, explained 24.6% of the variance in L2 reading performance among Slovenian school children.

However, L1 literacy-related difficulties do not provide a full explanation for poor L2 learning outcomes. For example, Alderson et al. (2015) found that 15% of weak readers in L2 English were strong readers in their L1 Finnish. Ferrari and Palladino's (2007) research with Italian children also demonstrated that L1 reading skills might not fully explain achievement in L2 learning. However, despite the fact that Kormos et al. (2019) found a relatively strong link between low level L1 skills and L2 reading, their results also revealed that only less than half of the students with official dyslexia identification belonged to the poor L2 reader group. One of the moderating factors that can explain variation in how strongly L1 literacy-related difficulties influence L2 reading is how L2 reading abilities are assessed.

2.3 Test fairness and time extension for candidates with SpLDs

The aim of ensuring fairness in assessment is to provide the opportunities for all test-takers to perform to the best of their knowledge. A key issue in language testing that is particularly relevant for the assessment of L2 reading skills of test-takers with SpLDs is the intricate relationship between test fairness and validity. A test is only valid if it does not systematically disadvantage any group of test-takers, and hence the construct of test validity needs to encompass test fairness (Kunnan, 2004).

Test fairness entails four important components: lack of bias, equitable treatment in the testing process, equality in the outcomes of testing, and fairness as an opportunity to learn (American Educational Research Association (AERA), 1999). First, a test is unbiased if it has no "construct irrelevant components that result in systematically lower or higher scores for identifiable groups of examinees" (AERA, 1999, p. 76). Both test content and response format can create systematic bias. Second, to ensure equitable treatment in the assessment process, the context and purpose of the assessment need to be considered. With regard to SpLDs, tests should be administered under testing conditions (e.g. time limits, physical environment) that allow candidates with disabilities to demonstrate the best of their knowledge. Third, equality in outcomes of testing can be ensured if there is evidence that no sub-groups of test-takers with similar levels of ability (e.g. those from different language or ethnic background) perform significantly differently on the tests. Finally, fairness as an opportunity to learn can be achieved by providing test-takers with equal opportunities for test preparation.

In some cases, when the accessibility of the test cannot be ensured by universal design principles, special arrangements are required to enable candidates with disabilities to display their knowledge. These special arrangements, which are called 'accommodations' or 'adjustments', can be categorised based on how they affect the validity of the test. Accommodations (e.g. extra time allocation) make adaptation to a test while still maintaining its validity. Modifications, such as allowing students to listen to a text while reading it, result in changes in test content or presentation that affect the construct validity of the test.

One type of accommodation that is frequently offered for students with SpLDs is extended time for completing the assessment tasks. Although the provision of more time might not impact on the construct validity of tests of reading and writing, research evidence in L1 literacy assessment suggests that it might also advantage students with no SpLDs (for a review see Sireci et al., 2005).

This means that the time limit for the test disadvantages students with SpLDs, but allowing them extended time might not necessarily give them an advantage. We also need to examine whether students with SpLDs benefit more from being given extra time than those with no SpLD; in other words, whether time extension gives them a differential boost. According to the *differential boost hypothesis* (Fuchs & Fuchs, 1999), making accommodations might be meaningful and might not pose a threat to the validity of the test, if it affects students with SpLDs differentially.

Proponents of time extension in exams argue that students with SpLDs take longer to read texts and formulate their response in writing, and hence should be given extra time so that they can achieve the maximum of their potential (e.g. Haladyna & Downing, 2004). However, the availability of more time, particularly in contexts where students need to provide longer answers to test questions, might overinflate test scores because it allows candidates with SpLDs to write more and give more detailed answers (Zuriff, 2000). A recent systematic overview of exams accommodations in all disciplines, including numeracy as well as literacy, by Duncan and Purcell (2019) found that out of the relevant 28 studies, 16 demonstrated that students with SpLDs were not unfairly advantaged by time extension, suggesting that time extension is a fair practice for both students with and without SpLDs. Six studies concluded that students with SpLDs gained from the additional time while test-takers with no SpLDs did not benefit from the time extension, which can be taken as evidence that extended time is fair for students with SpLDs. However, 12 studies revealed that the scores of test-takers with SpLDs were overinflated, indicating that time extension might be unfair to examinees with no SpLDs.

A meta-analysis by Gregg and Nelson (2012) focusing on the assessment of L1 literacy skills revealed that both adult and adolescent students with SpLDs scored significantly higher in tests of L1 reading skills when given extended time than did their peers with no SpLDs. Another analysis of the effect of time extension in L1 literacy and numeracy tests by Cahan, Nirel and Alkoby (2016), however, indicated that time extension policies can unfairly disadvantage non-disabled candidates who would also benefit from extra time. Their results revealed that approximately half of the non-disabled test-takers would have gained higher scores if they had been given extended time to complete the assessment. Some studies have also suggested that non-disabled students benefited more from time extension than their peers with SpLDs (e.g. Lewandowski, Cohen & Lovett, 2013; Lewandowski, Lovett & Rogers, 2008).

One of the possible reasons for contradictory findings is that the amount of extra time offered to test-takers in these studies has been different. In most contexts, students with SpLDs are granted either 50% or 100% additional time. In a recent overview of post-secondary exams in Canada, Sokal and Wilson (2017) found that in 70% of the cases, test-takers were awarded 50% more time. However, another study by Sokal and Vermette (2017) has revealed that 35% of university students who would have been entitled for extra time in exams did not use it at all and the average additional time used was only 17% (Sokal & Vermette, 2017). In the area of L1 literacy assessment, Spenceley, Wood, Valentino and Lewandowsky (2019) have recently shown that university students with SpLDs took 14% longer, on average, to complete a reading test in comparison with their peers with no SpLDs. Their results also revealed that the 20-minute allotted time for the reading comprehension test was not sufficient for either group of students. Test-takers with no SpLD required, on average, 15% more time, and those with SpLDs required 30% more time to answer all test items.

3. RESEARCH AIMS

As the above review shows, timing of test tasks can play an important role in how students perform in an exam, regardless of their SpLD status. Therefore, in order to design inclusive and universally accessible assessments, careful consideration needs to be given to how much time test-takers are allocated to complete a test task. Furthermore, it is also important that decisions about timing and granting time extensions to candidates with disabilities are made based on solid research evidence.

Promoting equality, diversity and inclusion has been among the key missions of the British Council. In the design process of the Aptis test, the British Council Assessment Research Group (ARG) has taken a proactive approach to inclusion and has paid particular attention to the accessibility of the test for diverse populations. As documented by Fairbairn and Spiby (2019), the ARG has aimed to remove barriers for test-takers with disabilities in the areas of communication with test-takers and test centres, test design, test production and scoring. Efforts of inclusive test creation and delivery have been recently assessed by case studies (cf. Fairbairn & Spiby, 2019), but no systematic investigation about the impact of time extension on test scores has yet been carried out. The case study in Chile presented by Fairbairn and Spiby also highlighted the issue of extra time because the standard procedure of granting additional 25% of time for the Aptis test was in conflict with the local assessment accommodations.

This study investigated the effect of different timing conditions on reading performance in the Aptis for Teens test. The study addressed the following research question.

RQ: Is there a differential effect of timing on different reading tasks in the Aptis for Teens test across students with varying levels of L1 skills?

Although information about potential reading-related disability was collected through an optional questionnaire item, disability status was not used as a primary independent variable in our research. On the one hand, this decision was taken because students might not disclose their disability. On the other hand, SpLDs are often misidentified in Hungary. More reliable information could be gained by assessing predictors of reading-related disabilities within the study. This also allowed us to treat L1 literacy skills as a continuous variable which reflects the dimensional nature of reading-related disabilities.

4. RESEARCH METHOD

4.1 Context and participants

Most upper primary schools in Hungary are state-funded and non-selective. Therefore, it was expected that the distribution of students with SpLDs would reflect the prevalence patters of SpLDs in the general population (approx. 10–15%). In order to sample from a wider range of contexts, two schools from different areas of Budapest, the capital city of Hungary, and one school from a smaller and economically less-developed town outside Budapest, were chosen.

The curricular objectives state that children should reach A2 level proficiency by the age of 14, but attainment varies across schools and there is additional variation among children depending on socio-economic status. Children generally start learning English in Year 2 and have three to four English classes per week. The level of proficiency of students in our research was between A1 and C1 (see Table 2 for reading test CEFR levels).

In total, complete data has been obtained from 100 students. Table 1 shows the number of students taking part in the research together with the gender distribution and their age. All students studied in Year 8, which is the final year of primary school. Since L1 literacy tests required that children have the same L1 background, all participants were native speakers of Hungarian. According to self-report data, 18 children had an official certificate of their SpLDs status.

Table 1: Students taking part in the research (n=100)

Site name	N	Gender		Age
		Male	Female	Years
School A	45	16	29	13.9
School B	24	9	15	13.8
School C	31	10	21	13.9
Total	**100**	**35**	**65**	**13.9**

Table 2: Students' reading proficiency level based on the Aptis for Teens test

Site name	A1	A2	B1	B2	C1
School A	1	21	12	10	1
School B	1	10	9	3	1
School C	4	15	10	1	1
Total	**6**	**46**	**31**	**14**	**3**

4.2 Instruments

Aptis for Teens Reading test

Students took the two versions of the Aptis for Teens Reading test in small groups in a computer lab. The test was administered in a counter-balanced design with each intact group starting with either the standard (45 minutes) or the extended (+25% = 56 minutes) version of the test.

Each test version contained 25 test items. In both test versions, the test items measured: Sentence Comprehension (SC) (5 items); Long Text Comprehension (LTC) (6 items); Short Text Comprehension (STC) (7 items); and Text Cohesion (TC) (7 items).

L1 reading comprehension test

To obtain a measure of participants' L1 reading comprehension skills, three texts were selected from a national test of reading comprehension developed by the Educational Research Institute of the Hungarian Ministry of Human Resources and Education in 2017. This test is designed by a team of experts in Hungary to assess the L1 literacy of primary school students in Year 8. Out of the three texts, two were informational and one narrative. One of the informational texts was about an everyday topic and the other about popular science. Table 3 gives an overview of the L1 reading comprehension texts.

Table 3: Texts and tasks in L1 reading comprehension test

	Entertaining news (informational-everyday)	Short-story (narrative)	Popular history (informational-scientific)
Number of items	13	9	9
MC items (4 options)	4	5	4
Short-answer items	5	4	4
True/false	4	0	0

Low-level L1 skills tests

To gather information about the participants' L1 word-level decoding skills and phonological awareness, we used a software called 3DM-H, which is an internationally recognised and nationally standardised computer-based assessment tool (Vaessen, Bertrand, Tóth, Csépe, Faísca, Reis & Blomert, 2010). 3DM-H is a Hungarian adaptation of the Dutch computerised cognitive test battery 3DM (Blomert & Vaessen, 2009). 3DM-H is a test intended for professional use and is accessible for academic research teams and institutions specialised in the diagnostics of SpLDs.

3DM-H has different subtests: 1. reading, 2. baseline response time, 3. orthography, 4. letter-speech sound identification, 5. phoneme deletion, 6. verbal working memory: phoneme span task, 7. visual-spatial memory: Corsi-cubes, 8. verbal working memory: syllable span task, 9. rapid automatic naming (RAN), 10. letter-speech sound discrimination, 11. visual-spatial memory: visual sequence. In Table 4 we describe the three selected sub-tests, which have been found to be the most reliable indicators of specific reading difficulties, i.e. dyslexia.

Table 4: 3DM-H sub-tests included in the study

Name of sub-test	Short description of the task	What it measures
Word reading	The student reads aloud lists of high-frequency words, low-frequency words and pseudowords presented on the computer screen within a set time (30 seconds/list)	The task assesses the fluency and accuracy of word-level decoding, where fluency is measured as the reading speed of correctly read items. The reading performance of high-frequency words reflects the automaticity of word recognition. The reading performance of pseudowords indicates the efficiency of word-level decoding processes that rely on mapping phonological representations on to orthographic form (letter–sound conversion).
Phoneme deletion	Pseudowords are presented auditorily without visual input. Students are instructed to delete given speech sounds (consonants) from different parts of the pseudowords (beginning, end, within a consonant cluster). (e.g. det without 't' [=de])	The task measures phonological awareness, i.e. the ability to manipulate sounds within words.
Rapid Automatic Naming (RAN)	Students name visually presented letters, numbers and pictures (e.g. fish, chair or dog) as fast as possible.	This task measures word naming speed, that is, the time needed to access and retrieve high frequency lexical items stored in the mental lexicon.

Personal background questionnaire

To establish a profile of the participant group, a brief personal background questionnaire was designed in the participants' L1. The questionnaire aimed to gather information on the participants' gender, age, year of study, language(s) spoken at home, residence abroad, SpLD status, length of learning English, and use of English outside the school context.

4.3 Procedures

Ethical approval was obtained from the Ethics Committee of Lancaster University. Parental consent was sought from all participants' caretakers and the children were also asked to express consent to participate in the study.

We contacted the English teachers of the participating students and provided them with the link for sample tasks of the Aptis for Teens Reading test so that they could practice the test tasks with their students. All teachers held a test preparation session with the students in which they familiarised them with the test.

The Aptis for Teens Reading test, the L1 reading comprehension test and the background questionnaire were piloted at the end of September 2018 with 28 students in School A.
We experienced no technical or content problems with any of the tests. The score report for piloting also showed a good spread of scores from A1 to C1 level in the Aptis for Teens Reading test and a normal score distribution for the L1 reading comprehension test. The three sub-tests of 3DM-H intended to assess L1 word-level decoding, rapid automated naming and phonological awareness also functioned well in the pilot. An inspection of the pilot study data for the three sub-tests of 3DM-H showed normal distribution.

As originally planned, students took the two versions of the Aptis for Teens Reading test in small groups in the school computer labs. The tests were administered in a counter-balanced order with each group starting with either the standard (30 minutes) or the extended (+25% = 37.5 minutes) version of the test (see Table 5 below) with a 15-minute break between the two parts of the reading tests. Half of the participants in each group received Version A and the other half Version B of the test. After another break, students completed the paper-based Hungarian L1 reading comprehension test (45 minutes) and filled in the participant background questionnaire (5 minutes).

Table 5: The procedures of the group-testing phase of the study

1st Aptis test group session	2nd Aptis test group session	3rd group session
Standard reading version A	Extended reading version B	L1 reading comprehension test
Standard reading version B	Extended reading version A	L1 reading comprehension test
Extended reading version A	Standard reading version B	L1 reading comprehension test
Extended reading version B	Standard reading version A	L1 reading comprehension test

The three sub-tests of 3DM-H were administered individually in a quiet room of the school on separate days from the group testing sessions. The students completed these tests within 15 minutes in an individual session with the trained research assistant.

5. RESULTS

5.1 Descriptive statistics and preliminary analyses

First, we examined the distribution of the scores of the low level L1 skills and L1 reading comprehension tests. Except for the phoneme deletion task, the distribution was fairly symmetric and the kurtosis values were also within the -1.96 and +1.96 range (see Table 6). The Cronbach alpha reliability of the L1 reading comprehension test was .891.

Table 6: The descriptive statistics of low level L1 skills and L1 reading comprehension tests

	Minimum	Maximum	Mean	SD	Skewness	Kurtosis
Frequent word reading fluency	2.96	14.00	7.59	2.10	.222	-.068
Rare word reading fluency	2.25	11.05	6.26	1.77	.209	.366
Non-word reading fluency	2.02	7.35	4.69	1.16	-.164	-.269
Phoneme deletion	1.11	3.49	.64	.95	.793	.392
RAN letter	.72	3.23	1.86	.45	.084	.375
RAN number	1.66	3.20	2.47	.35	-.116	-.614
RAN objects	1.08	2.06	1.60	.21	.072	-.404
L1 reading comprehension	8.00	31.00	21.56	5.25	-.487	-.266

(RAN – rapid automated naming)

Next, we examined the correlations between these variables. As can be seen in Table 7, all variables were significantly correlated. Most correlations were moderate, and the relationship between phoneme deletion speed and L1 reading comprehension was weak. Strong correlations were found among the L1 word-level reading measures. Phoneme deletion speed also strongly correlated with the reading fluency of frequent words. L1 text comprehension was strongly associated with rare and non-word reading fluency.

Table 7: The correlations of low level L1 skills and L1 reading comprehension tests

	RWRF	NWRF	PD	RANl	RANn	RANo	L1 RC^
Frequent word reading fluency (FRWF)	.689**	.631**	.502**	.609**	.377**	.367***	.368**
Rare word reading fluency (RWRF)		.848**	.458**	.680**	.446**	.357**	.534**
Non-word reading fluency (NWRF)			.408**	.646**	.429**	.388**	.528**
Phoneme deletion (PD)				.414**	.315**	.308**	.211*
RAN letter (RANl)					.375**	.503**	.305**
RAN number (RANn)						.301**	.354**
RAN object (RANo)							.405**

^: L1 RC: L1 reading comprehension
* $p < .05$ ** $p < .001$

In order to reduce the number of variables for further analyses, we conducted factor analyses. The initial factor analysis showed that the rapid automated naming of objects (.367) and letters (.338) had a relatively low level of commonality with the other variables. We then conducted another factor analysis with the remaining variables. Kaiser-Meyer-Olkin value of .862 exceeded the recommended .6 and Bartlett's Test of Sphericity reached statistical significance ($p < .001$) supporting the factorability of the correlation matrix. The screeplot revealed a clear break after the first component. One factor was extracted (Eigenvalue 3.669) which explained 61.65% of the variance. Based on the results of the factor analysis, we deemed it appropriate to create a composite score using regression factor scores (Tabachnick & Fidell, 2001).

We also wanted to check the validity of our L1 measurements by examining whether students with an official certificate of their SpLDs differed along the low-level L1 skill variables and the L1 reading comprehension scores. The results presented in Table 8 showed that there was a significant difference with a large effect size in all the examined variables including the regression factor score.

Table 8: Differences between students with (n= 18) and without SpLDs (n=82) in terms of low-level L1 and L1 comprehension skills

	SpLD	Mean	Std. deviation	*t*	Cohen's *d*
Frequent word reading fluency	no	8.01	1.93	4.93**	1.33
	yes	5.57	1.71		
Rare word reading fluency	no	6.63	1.61	5.30**	1.43
	yes	4.47	1.39		
Non-word reading fluency	no	4.94	1.05	5.38**	1.48
	yes	3.49	.89		
Phoneme deletion	no	.83	.89	4.96**	1.36
	yes	-.27	.71		
RAN number	no	2.53	.32	4.64**	1.32
	yes	2.14	.33		
L1 reading comprehension	no	22.80	4.44	6.65**	1.37
	yes	16.15	5.20		
Regression factor score	no	.26	.83	7.12**	1.84
	yes	-1.25	.75		

**$p<.001$

We also measured the time it took the students to complete the tests using a stopwatch. Although this method was not entirely accurate, the data showed that students completed the test on average within 19 minutes, which was around 35% less of the original time allocated (30 minutes). Only five students went over 30 minutes in answering the test items in the extended time version, and none of these participants had an official certificate of their SpLDs. There was no significant difference between students with SpLDs and without SpLDs in terms of the time they needed to finish the test $t(98) = 1.26$, $p = .208$.

5.2 Differential effect of timing on different reading tasks across students with varying levels of L1 skills

Our analysis aimed to answer the overarching research question: *Is there a differential effect of timing on different reading tasks across students with varying levels of L1 skills?* To answer this research question, 100 students completed two versions of the Aptis for Teens reading test (A and B). Each test version contained 25 test items, but these test items differed between test versions in their content. Thus, test items were nested within the different test versions (A vs. B)[1].

[1] There was no statistically significant difference between scores on the A and B versions of the Aptis for Teens reading test either in the short (t (99) = .362 p = .718) or the long version (t (99) = .956 p = .341)

In both test versions, the test items measured: Sentence Comprehension (SC); Long Text Comprehension (LTC); Short Text Comprehension (STC); and Text Cohesion (TC). The test versions were randomised so that students did either short version of the Aptis test A and then long version of the test B, or short version A and then long version B. Overall, we analysed 5,001 observations from 100 students. The Cronbach alpha reliability of the standard-timing test version A was .855, for standard timing test Version B .833, for extended timing version A .871 and for extended timing version B .836.

Table 9: Descriptive statistics: Mean probabilities; Tasks across Test Time, at average L1 Skills

Task	Test time	Mean	Standard deviation
SC	Short	.83	.37
	Long	.81	.40
LTC	Short	.22	.41
	Long	.20	.40
STC	Short	.46	.50
	Long	.48	.50
TC	Short	.54	.50
	Long	.50	.50

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

The descriptive statistics table (Table 9) shows that students had the highest probability of getting a question right on the SC task of the Aptis for Teens test. Conversely, students had the lowest probability of getting a comprehension question right in the LTC task of the Aptis for Teens test. Table 9 also shows that the mean differences between different test times were relatively small for each task.

To examine the factors that influenced the log-odds of reading comprehension accuracy on the Aptis for Teens test, we used Generalised Linear Mixed-Effects Models (GLMMs). We built these models using the glmer function in the lme4 package (Bates, Mächler, Bolker & Walker, 2015) in R (R Core Team, 2019). GLMMs were theoretically appropriate for this analysis, because we had item-level accuracy data that followed a binomial distribution. In other words, for each question the only possible outcome was either a correct response or an incorrect response. Thus, we had to model the probability of getting a comprehension question right, and GLMMs allowed us to do that.

The fixed effects included: Test Time (Short vs. Long); Task (SC, LTC, STC, and TC); and a principal component of L1 Skill. To make interactions more interpretable, the L1 Skills variable was centred at the mean. It was also scaled by two standard deviations in order to guard against understating the importance of L1 skills in reading comprehension accuracy (Gelman, 2008). To minimise the Type I error rate of our predictions, our models considered random variation between participants and test items (Jaeger, 2008). Specifically, we fit our models with random effects to account for variation in by-individual and by-item-nested-within-test-version accuracy (random intercepts). We also tried to account for variation in the strength of the effects of predictor variables on reading comprehension accuracy, by fitting slopes of the fixed effects (random slopes) associated with the differences between participants and the test items (Baayen, Davidson, & Bates, 2008).

In the first step of the analysis, we established whether the addition of fixed effects and interactions, while keeping the random effects constant, improved the extent to which the observed data matched the values expected by theory, also referred to as the model's goodness-of-fit. We used the Likelihood Ratio Test (LRT; Baayen, 2008) comparisons to compare simpler models with the more complex ones. We describe the results of the LRT comparisons, but report estimates of the final model only.

We progressed through a series of models, starting with a minimal model of the log-odds reading accuracy, with the random effects of students and items nested within test versions on intercepts (Model 1). The minimal model was compared to a model with the fixed effects of: Test Time, Task, and L1 Skills (Model 2). The LRT revealed that adding complexity to the model was justified. Model 2 provided a better fit to the data than Model 1 $\chi^2(5) = 95.42$, $p < .001$. We compared the model with the main effects to a model with added interactions of: Test Time by Task; Test Time by L1 Skills; Test Time by Task by L1 Skills (Model 3). Increasing model complexity further improved the model fit, $\chi^2(10) = 32.50$, $p < .001$. On the grounds of improvement in the goodness-of-fit, and to answer our research questions, we decided to keep the two-way and three-way interactions in our final model.

In the second step of the analysis, we evaluated whether the inclusion of all random intercepts was necessary using pairwise LRT comparisons of models with stable fixed effects, but with a varying random effects structure. We compared: (i) Model 3 with the random effects of students and items nested within test versions on intercepts; (ii) Model 3 with random effects of students on intercepts; (iii) Model 3 with random effects of items nested within test versions on intercepts. The LRT revealed that both random intercepts improved the model fit. There were significant differences between Models (i) and (ii) ($\chi^2(1) = 107.99$, $p < .001$), and Models (i) and (iii) ($\chi^2(1) = 645.29$, $p < .001$). Thus, the inclusion of random effects of students and items nested within test versions on intercepts was justified.

Next, as recommended by Barr, Levy, Scheepers and Tily (2013), we fit our model with random slopes, random differences between students and items nested within test versions, in the slopes of the fixed effects due to Test Time, Task, and L1 Skills. This was motivated by the wish to further minimise the Type I error rate. We found that a Maximal Likelihood Model, consisting of terms corresponding to random effects of students and items nested within test versions on the slopes of Test Time, Task, and L1 Skills failed to converge. Thus, following the relatively recent recommendations in statistical science to keep the model maximal within the boundaries of what the data can support (Bates, Kliegl, Vasishth, & Baayen, 2018; Matuschek, Kliegl, Vasishth, Baayen & Bates, 2017), we established the utility of random slopes using the LRT.

We found that only the random slopes of Test Time on the random intercept of students, significantly improved the model fit ($\chi^2(2) = 9.34$, $p = .009$) and did not lead to over-fitting. Attempts to fit other random slopes to the model led to either non-convergence or singular autocorrelations. Singular autocorrelations are indicative of over-fitting, suggesting that a model is too complex for the information provided by the study's data (Bates et al., 2018). Simulations show that models lose power to find real effects, Type II error, if their complexity is not supported by the data (e.g. Matuschek et al., 2017). Consequently, our final model contained random slopes of Test Time on random intercept of participants only. We show the code used to fit the final model below:

*Reading Comprehension Accuracy ~ Test Time * Task * L1 Skills +*
(Test Time + 1|Participant) + (1|Test Version:Item)

The final model accounted for 50.60% of the variance associated with reading comprehension accuracy. The random effects accounted for the majority of the variance (27.32%), showing that a lot of variation in individuals' comprehension accuracy was due to random differences between participants and test items within and between test versions. The rest of the variance in reading comprehension accuracy (23.28%) was accounted for by the predictor variables, indicating that a substantial amount of variation in reading comprehension accuracy was predicted by the effects of Test Time, Task, and L1 Skills.

We report a summary of the final model in Table 10 where we supplement the log-odds estimates with Odds Ratio (OR) estimates. It is important to mention that the summary table of the final model should not be interpreted directly, as all the coefficients are estimated against reference level categories. For example, the significant coefficient for Task LTC predicts that students were 50 (1/.02) times less likely to get a comprehension question right for Task LTC than in Task SC, but only for reference level of Test Time (Short) and keeping L1 skills at average.

For a more intuitive interpretation of estimates, subsequent multiple comparisons tables should be considered (Tables 11–13), where we adjusted the *p*-value for multiple comparisons using normal approximation. Figures 1 and 2 supplement the multiple comparisons tables.

Table 10: Summary of the Reading Model

Fixed effects	Estimate	Odds ratio	Standard error	*z*-value	*p*
(Intercept)	2.19	8.91	.25	8.87	***
Test Time: Long	-.28	.75	.20	-1.41	.16
Task: LTC	-3.88	.02	.29	-13.57	***
Task: STC	-2.36	.09	.27	-8.64	***
Task: TC	-1.92	.15	.28	-6.85	***
L1 Skills	1.32	3.74	.35	3.78	***
Test Time: Long by Task: LTC	.23	1.25	.25	.89	.37
Test Time: Long by Task: STC	.40	1.49	.23	1.75	.08
Test Time: Long by Task: TC	.07	1.08	.23	.31	.75
Test Time: Long by L1 Skills	-.31	.73	.37	-.84	.40
Task: LTC by L1 Skills	-1.51	.22	.35	-4.34	***
Task: STC by L1 Skills	-.62	.54	.32	-1.95	.05
Task: TC by L1 Skills	-.76	.47	.32	-2.35	*
Test Time: Long by Task: LTC by L1 Skills	.83	2.29	.48	1.73	.08
Test Time: Long by Task: STC by L1 Skills	-.03	.97	.43	-.06	.95
Test Time: Long by Task: TC by L1 Skills	.71	2.03	.45	1.59	.11
Random Effects (Intercepts)	Random Slopes	Variance	Standard Deviation	Correlation	
Participants		1.30	1.14		
	Test Time: Long	.25	.50	-.26	
Item (nested within Test Version)		.27	.52		

Note 1. * = $p < .05$; ** = $p < .01$; *** = $p < .001$. *Note 2.* Short is the reference level for Test Time; Task SC is the reference level for Task. *Note 3.* L1 Skills is centred and standardised, by two standard deviations, factor score.

To look into the role of Test Time and Task in L2 reading accuracy performances, we investigated how odds of getting a question right varied across Test Time and L1 skills. We found significant differences in comprehension accuracy log-odds between Tasks within each Test Time (Table 11). In each comparison, students were more likely to answer comprehension questions correctly in Task SC versus any other task. One observation that can be made is that the predicted differences between the tasks were smaller for students with below average L1 skills compared to those with average and above average L1 skills.

Table 11: Multiple comparisons: Differences between Tasks across Test Time, varying L1 Skill level

Comparison	Estimate	Odds ratio	Standard error	*z*-value	*p*
Short (L1 Skills at average)					
LTC - SC	-3.88	.02	.29	-13.57	***
STC - SC	-2.36	.09	.27	-8.64	***
TC - SC	-1.92	.15	.28	-6.85	***
STC - LTC	1.52	4.58	.24	6.31	***
TC - LTC	1.96	7.10	.25	7.79	***
TC - STC	.44	1.55	.24	1.82	.26
Long (L1 Skills at average)					
LTC - SC	-3.65	.03	.28	-13.12	***
STC - SC	-1.96	.14	.27	-7.38	***
TC - SC	-1.85	.16	.27	-6.74	***
STC - LTC	1.69	5.43	.24	7.04	***
TC - LTC	1.81	6.08	.25	7.22	***
TC - STC	.11	1.12	.24	.47	.97
Short (L1 Skills above average)					
LTC - SC	-5.39	.005	.48	-11.17	***
STC - SC	-2.98	.05	.44	-6.69	***
TC - SC	-2.68	.07	.45	-5.90	***
STC - LTC	2.41	11.17	.38	6.42	***
TC - LTC	2.71	15.06	.39	6.94	***
TC - STC	.30	1.35	.35	.85	.83
Long (L1 Skills above average)					
LTC - SC	-4.34	.01	.45	-9.66	***
STC - SC	-2.61	.07	.42	-6.25	***
TC - SC	-1.90	.15	.43	-4.42	***
STC - LTC	1.73	5.64	.37	4.74	***
TC - LTC	2.44	11.48	.38	6.39	***
TC - STC	.71	2.03	.35	2.03	.18
Short (L1 Skills below average)					
LTC - SC	-2.37	.09	.42	-5.70	***
STC - SC	-1.74	.18	.39	-4.45	***
TC - SC	-1.16	.31	.40	-2.91	*
STC - LTC	.63	1.88	.37	1.72	.31
TC - LTC	1.21	3.34	.38	3.19	**
TC - STC	.58	1.78	.35	1.64	.35
Long (L1 Skills below average)					
LTC - SC	-2.97	.05	.42	-7.07	***
STC - SC	-1.32	.27	.38	-3.45	**
TC - SC	-1.80	.17	.40	-4.50	***
STC - LTC	1.65	5.22	.37	4.46	***
TC - LTC	1.17	3.22	.39	3.02	*
TC - STC	-.48	.62	.35	-1.37	.52

Note 1. * = $p < .05$; ** = $p < .01$; *** = $p < .001$. *Note 2.* Average is at mean L1 Skills; above average is two standard deviations above mean L1 Skills; below average is two standard deviations below mean L1 Skills.

Critically, we found that time extension did not predict comprehension accuracy of students of any L1 skills (Table 12; Figure 1; Figure 2). Table 12 illustrates that students with different L1 skills levels were not predicted to perform significantly better or worse for any of the tasks with time extension. Figure 1 shows that the 95% confidence intervals associated with the two levels of Test Time (Short vs. Long) were overlapping for students of every L1 skills level and across every task. Figure 2 focuses on the comparisons between the two Test Time conditions at average, above average, and below average L1 skills only. Overall, our model provides evidence to suggest that, given our data, there does not seem to be a differential effect of timing on different reading tasks across students with varying levels of L1 skills.

Table 12: Multiple comparisons: Test Time by Task; varying L1 Skills

Comparison	Estimate	Odds ratio	Standard error	*z*-value	*p*
L1 Skills at average					
Long SC - Short SC	-.28	.75	.20	-1.41	.49
Long LTC - Short LTC	-.05	.95	.16	-.34	.995
Long STC - Short STC	.12	1.12	.13	.87	.85
Long TC - Short TC	-.21	.81	.14	-1.46	.46
L1 Skills above average					
Long SC - Short SC	-.59	.55	.46	-1.28	.59
Long LTC - Short LTC	.46	1.58	.36	1.26	.60
Long STC - Short STC	-.22	.80	.30	-.75	.91
Long TC - Short TC	.19	1.21	.32	.59	.96
L1 Skills below average					
Long SC - Short SC	.03	1.03	.38	.08	1.00
Long LTC - Short LTC	-.57	.57	.36	-1.56	.39
Long STC - Short STC	.45	1.58	.30	1.53	.41
Long TC - Short TC	-.61	.55	.32	-1.89	.21

Note. Average is at mean L1 Skills; above average is two standard deviations above mean L1 Skills; below average is two standard deviations below mean L1 Skills.

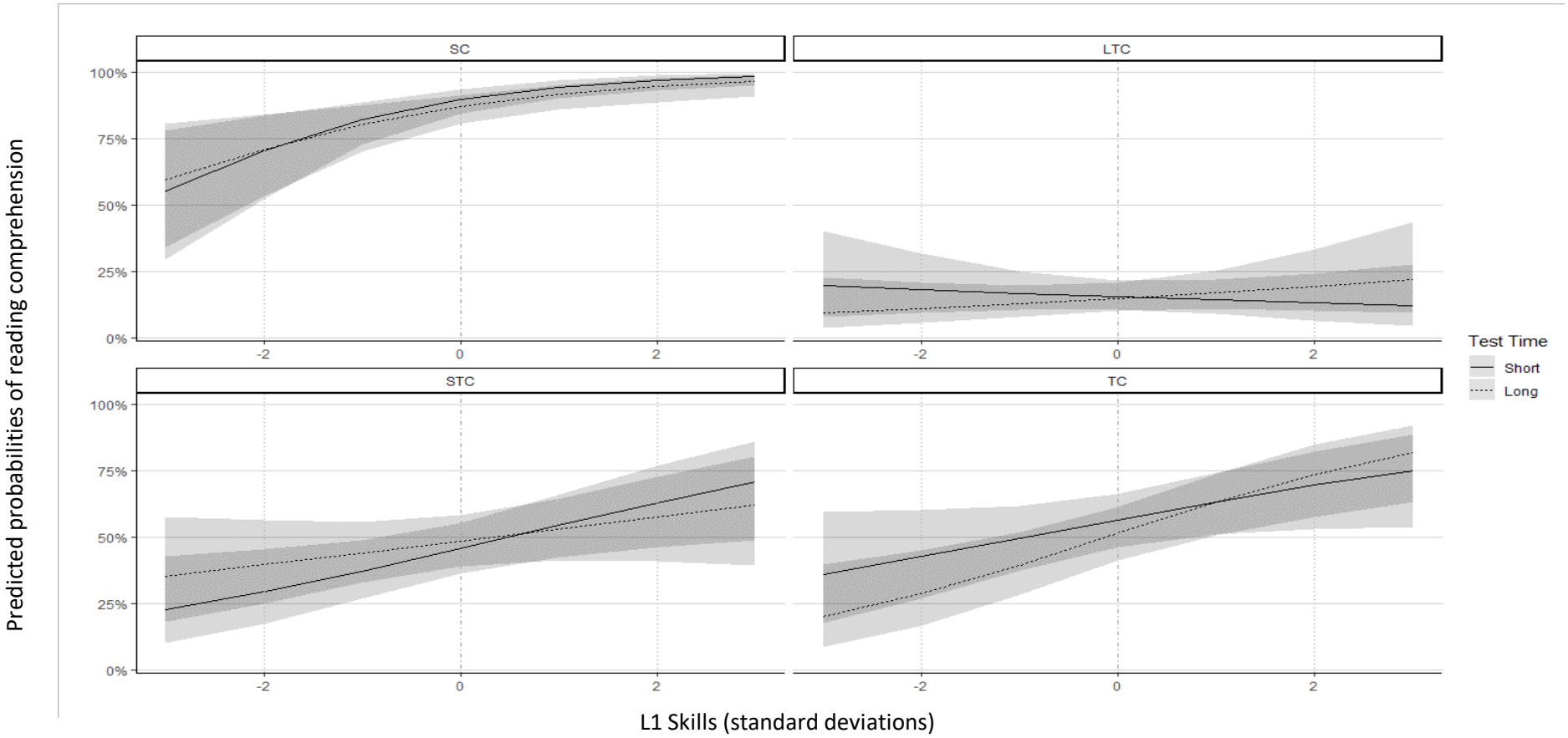

Figure 1: The effects of L1 Skill variation on reading comprehension accuracy across Tasks and Test Time.

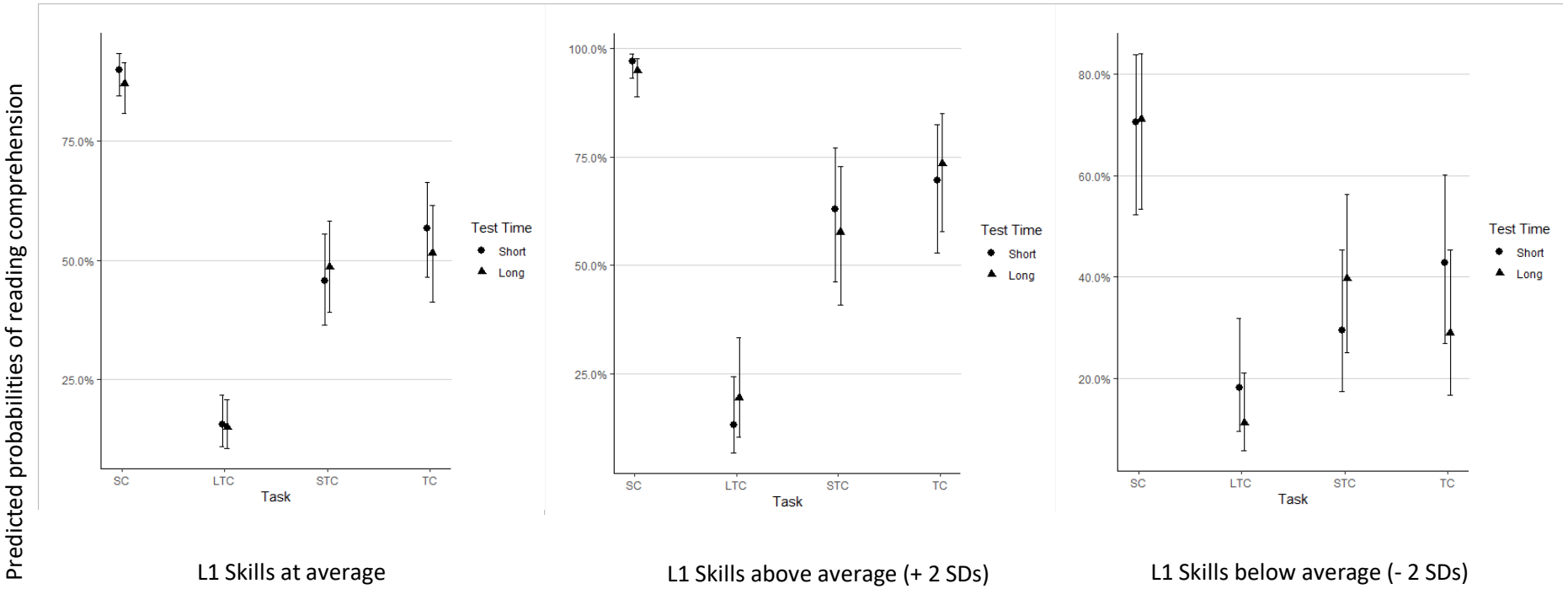

Figure 2: The effects of Task and Test Time on reading comprehension accuracy, varying L1 Skills

Our comparisons also revealed that students with higher L1 skills were more likely than those with average L1 skills likely to answer comprehension questions correctly related to SC (3.74 times) and STC (2.02 times) tasks without time extension (Table 13). Under time extension condition (Long), students with higher L1 skills were 2.74 and 2.61 times more likely than students with average L1 skills to answer comprehension questions correctly related to SC and TC respectively. Overall, in all cases but LTC without time extension, an increase in L1 skills was associated with an increase in reading comprehension accuracy, although this increase was not always significant. Reading comprehension accuracy was most robustly predicted by an increase in L1 skills on the SC task, whereby those with higher L1 skills were more likely to answer comprehension questions correctly under both Test Time conditions.

Table 13: Multiple comparisons: The effect of an increase in L1 Skills (2 SDs) by Test Time across Tasks

Comparison	Estimate	Odds ratio	Standard error	*z*-value	*p*
Short					
SC	1.32	3.74	.35	3.78	***
LTC	-.19	.83	.32	-.60	.93
STC	.70	2.02	.29	2.44	*
TC	.56	1.75	.30	1.90	.17
Long					
SC	1.01	2.74	.33	3.04	**
LTC	.32	1.38	.31	1.03	.67
STC	.36	1.44	.28	1.29	.49
TC	.96	2.61	.30	3.25	**

Note 1. * = $p < .05$; ** = $p < .01$; *** = $p < .001$.

6. DISCUSSION AND RECOMMENDATIONS

6.1 The effect of test time

In our research we were interested in the differential effect of timing on Aptis for Teens reading test tasks across students with varying levels of L1 skills. A preliminary analysis of the test timing revealed that the typical allotted time was sufficient for 95% of the participants and none of the participants who exceeded the time limit had an officially certified SpLD. This result suggests that the reading component of the Aptis for Teens test meets the criteria of universal accessibility in terms of its timing. This analysis has also shown that the established time limit is about 35% longer than students on average needed to complete the test tasks. If we take into account that the maximum time a student required to finish the test was 35 minutes, we can conclude that if we add 50% time to an average test completion time, it might allow every student to display the best of their abilities. Therefore, in future test development and piloting phases, it is important to carefully examine how long students on average take to complete a reading test, and it is recommended to add a 50% margin to accommodate slower test-takers.

The results of the mixed-effects modelling also confirm the findings regarding test timing. Our statistical model revealed no significant effect of the time extension and no significant interaction effects between time extension, L1 skills and test tasks. This suggests that time extension did not boost students' scores and did not confer an advantage for students with low-level L1 skills either. This finding is different from most other research previously conducted in the field of L1 literacy assessment (cf. Duncan & Purcell, 2019; Gregg & Nelson, 2012) which either found that time extension was helpful for all students regardless of SpLD status or that candidates with SpLDs benefited from the extra time. One of the reasons for the difference in findings might be that most L1 studies used strictly timed tests and none of them had a careful counter-balanced research design. Furthermore, none of the previous studies applied mixed-effects modelling in their analyses, which means that random variation in their data could have significantly influenced the findings. Moreover, most previous research either compared the performance of students with SpLDs under extended time conditions with that of test-takers with no SpLDs under standard conditions, or examined the scores of these two groups on extended versions.

Our results confirm the importance of universal test design and allowing a generous margin of around 50% extra time from the mean completion time for all test-takers. Similar recommendation about ensuring that sufficient time is available for all candidates were also made by Cahan, Nirel and Alkoby (2016), whose study indicated that time extension policies can disadvantage non-disabled candidates who might also perform better under extended conditions. This argument is also underscored by our observation that the candidates who used up some of the available extra time in the extended-timing version did not have an official SpLD diagnosis. In fact only one of the five students who needed additional time had below average L1 skills (by approximately 1 SD).

Our analysis has also demonstrated that time extension was unlikely to benefit students' performance on any of the tasks, including even the comprehension of a longer passage. It is important to note, however, that the longest text students had to read contained about 300–350 words. Most previous research in the field of L1 literacy had been conducted with older students and longer and more complex texts. The exams investigated also tended to be longer and contained more reading passages. Therefore, another reason for the lack of time effects might be due to the nature of the tasks and the length of texts in the Aptis for Teens test. Furthermore, the Aptis for Teens reading test is computer administered and students need to click on the right answers. Most previous studies in the L1 field used paper-based tests and also contained items that required written answers of varying length.

6.2 Task effects

The possible impact of test tasks on the generalisability of our findings can also be supported by the data on task difficulty and the comparison of students' scores on different tasks. As is apparent from the descriptive statistics (see Table 9), as well as the multiple comparisons (Table 11), students were more likely, and were predicted to be more likely, to answer comprehension questions correctly on the sentence comprehension task and less likely to do so on the long text comprehension task. In fact, the facility values for both of these tasks are outside the acceptable range of .3 to .7 (Farhady, 2012), suggesting that the sentence comprehension task was too easy and the long text comprehension task too difficult for our participants. The multiple comparisons across L1 skill levels also confirm this conclusion. In all conditions except for the standard-length version for the below average L1 skill students, the probability of answering questions correctly on the sentence comprehension task was predicted to be significantly higher compared to any other task. Furthermore, the probability of providing correct answers on the long text comprehension was predicted to be significantly lower than on the other tasks regardless of L1 skill level. Under the standard timing condition, participants with below average L1 skills performed similarly on short and long-text comprehension. Answering questions that required students to understand cohesive relations within the text and understanding a short text was predicted to be equally and moderately difficult for our sample.

Based on these results, we believe that further analyses on other test populations are needed to confirm the psychometric properties of the sentence comprehension and long text comprehension tasks for teenage L2 learners. In our investigated context, these two tasks might not give accurate insights into students' L2 reading skills. The task effects might also need future attention as our mixed-effects modelling suggests that a large percentage of test scores in our data set was due to random differences between participants and test items within and between test versions. These random variations, as well as potential ceiling and floor effects in two of the four tasks, might also explain why we did not detect a significant impact of extended time in our research.

6.3 The role of low-level L1 skills

Our mixed-effects modelling predicted that low-level L1 skills combined with L1 reading comprehension scores impact on L2 reading performance, which is in line with previous studies in the field that have demonstrated a link between L1 and L2 reading performance (e.g. Alderson et al., 2015; Erdos et al., 2014; Kormos et al., 2019; van Gelderen et al., 2004). In this regard, our results lend support to the the linguistic interdependence hypothesis (Cummins, 1979) and Geva and Ryan's (1993) common underlying processes framework, which argue that both L1 and L2 reading achievement is influenced by a key set of individual difference factors.

However, it is worth noting that the impact of L1 skills was found to vary depending on the reading task. In the standard timing condition, students with above average L1 skills were predicted to score higher on the sentence and short-text comprehension tasks, but not on tasks that involved reading a longer text and understanding the cohesive links within a text. In the longer test administration condition, the impact of L1 skills was predicted to remain stable on the sentence comprehension task, but instead of the short-text comprehension task, it predicted performance on the cohesion task. As low-level L1 skills were also included in the composite L1 skills factor score, it is not unexpected that these skills contributed to L2 sentence-level comprehension. Sentence-level comprehension requires the deployment of lower-level reading processes of fast and automatic word recognition, which was assessed by our low-level L1 skills tests, and lexico-syntactic processing (Perfetti, 2007). For understanding isolated sentences, readers do not tend to rely on higher-level processes such as comprehension monitoring, inference making, prior knowledge, and standards of coherence (Grabe, 2014). These higher-level processes also play a less important role in understanding shorter texts, which might explain why low-level L1 skills were found to predict short-text comprehension.

The understanding of longer text seems to be less strongly influenced by L1 skills probably because adolescents at this age can deploy metacognitive strategies to compensate for potential weaknesses in low-level decoding skills (cf. Van Gelderen et al., 2004).

Our findings, that demonstrate the variation in the influence of L1-related factors depending on tasks and test administration condition, are in line with Jeon and Yamashita's (2014) meta-analysis, which showed that task-related variables also influence L2 reading performance. Furthermore, they underscore the importance of investigating how the effects of individual differences, in interaction with the effects of text features, predict reading comprehension (cf. Kulesz, Francis, Barnes & Fletcher, 2016). Understanding the effects of interactions of individual difference factors and text and task characteristics is crucial if test designers want to ensure that students with particular characteristics are not disadvantaged by a given test task. Uncovering the effects of these interactions can also provide useful validity evidence because, if one particular individual characteristic has an unexpectedly strong impact on performance in a test task, the task might assess a skill that is potentially unrelated to L2 competence.

In our research, the fact that the impact of L1 skills was not entirely stable across the standard and extended timing conditions indicates that somewhat different processes might have been involved in how students solved the tasks with, and without, time constraints. This again speaks for the importance of establishing a universally accessible time-limit that allows all test-takers, regardless of their SpLD status and their L1 skills, to perform to the best of their ability.

REFERENCES

AERA/APA/NCME. (1999*). Standards for Educational and Psychological Testing*. Washington, DC: Author.

Alderson, J.C. (1984). Reading in a foreign language: A reading problem or a language problem. In J. C. Alderson & A. H. Urquhart (Eds.), *Reading in a foreign language* (pp. 1–24). London: Longman.

Alderson, J.C., Haapakangas, E-L., Huhta, A., Nieminen, L., & Ullakonoja, R. (2015). *Diagnosing reading in a second or foreign language*. London: Routledge.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). Washington, DC: Author.

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

Baayen, R.H., Davidson, D.J., & Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390–412.

Barr, D.J., Levy, R., Scheepers, C., & Tily, H.J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68, 255–278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). *Parsimonious mixed models*. arXiv:1506.04967v2.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48.

Bernhardt, E.B., & Kamil, M.L. (1995). Interpreting relationships between L1 and L2 reading: Consolidating the linguistic threshold and the linguistic interdependence hypotheses. *Applied Linguistics*, 16(1), 15–34.

Blomert, L. & Vaessen, A. (2009). *Differentiaal Diagnostiek van Dyslexie: Cognitieve analyse van lezen en spellen.* Boom Test, Amsterdam, NL.

Cahan, S., Nirel, R., & Alkoby, M. (2016). The extra-examination time granting policy: A reconceptualization. *Journal of Psychoeducational Assessment*, 34(5), 461–472. doi: 10.1177/0734282915616537

Crombie, M. (1997). The effects of specific learning difficulties (dyslexia) on the learning of a foreign language at school. *Dyslexia*, 3, 27–47.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, 49, 222–251.

Duncan, H., & Purcell, C. (2019). Consensus or contradiction? A review of the current research into the impact of granting extra time in exams to students with specific learning difficulties (SpLD). *Journal of Further and Higher Education*, DOI:/abs/10.1080/0309877X.2019.1578341

Erdos, C., Genesee, F., Savage, R., & Haigh, C. (2014). Predicting risk for oral and written language learning difficulties in students educated in a second language. *Applied Psycholinguistics*, 35(2), 371–398.

Fairbairn, J., & Spiby, R. (2019). Towards a framework of inclusion: developing accessibility in tests at the British Council. *European Journal of Special Needs Education*, 34, 236–255.

Farhady, H. (2012). Principles of language assessment. In C. Coombe & B. O'Sullivan (Eds), *The Cambridge Guide to Second Language Assessment* (pp. 37–46). Cambridge University Press.

Ferrari, M., & Palladino, P. (2007). Foreign language learning difficulties in Italian children: Are they associated with other learning difficulties? *Journal of Learning Disabilities*, 40, 256–269.

Fuchs, L.S., & Fuchs, D. (1999). Fair and unfair testing accommodations. *School Administrator*, 56, 24–29.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27(15), 2865–2873.

Geva, E., & Ryan, E.B. (1993). Linguistic and cognitive correlates of academic skills in first and second languages. *Language Learning*, 43, 5–42.

Geva, E., & Wiener, J. (2014). *Psychological assessment of culturally and linguistically diverse children and adolescents: A practitioner's guide*. New York: Springer Publishing Company.

Geva, E., Wade-Woolley, L., & Shany, M. (1993). The concurrent development of spelling and decoding in two different orthographies. *Journal of Literacy Research*, 25, 383–406.

Gough, P.B., & Tunmer, W.E. (1986). Decoding, reading and reading disability. *Remedial and Special Education*, 7, 6–10.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice.* Cambridge: Cambridge University Press.

Gregg, N., & Nelson, J.M. (2012). Meta-analysis on the effectiveness of extra time as a test accommodation for transitioning adolescents with learning disabilities: More questions than answers. *Journal of Learning Disabilities*, 45(2), 128–138. doi: 10.1177/0022219409355484

Haladyna, T.M., and S.M. Downing. (2004.) Construct-Irrelevant Variance in High-Stakes Testing. *Educational Measurement: Issues and Practice*, 23 (1): 17–27.

Helland, T., & Kaasa, R. (2005). Dyslexia in English as a second language. *Dyslexia*, 11, 41–60.

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.

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

Kirby, J.R., Georgiou, G.K., Martinussen, R., & Parrila, R. (2010). Naming speed and reading: From prediction to instruction. *Reading Research Quarterly*, 45, 341–362.

Kormos, J. (2017). *The second language learning processes of students with specific learning difficulties*. New York: Routledge.

Kormos, J., & Mikó, A. (2010). Diszlexia és az idegen-nyelvtanulás folyamata [Dyslexia and the process of second language acquisition]. In J. Kormos and K. Csizér (Eds.), *Idegennyelv-elsajátítás és részképességzavarok. [Foreign language acquisition and learning disabilities]* (pp. 49–76). Budapest: Eötvös Kiadó.

Kormos, J., Babuder, M.K., & Pižorn, K. (2019). The role of low-level first language skills in second language reading, reading-while-listening and listening performance: A study of young dyslexic and non-dyslexic language learners. *Applied Linguistics*.

Kulesz, P.A., Francis, D.J., Barnes, M.A., & Fletcher, J.M. (2016). The influence of properties of the test and their interactions with reader characteristics on reading comprehension: An explanatory item response study. *Journal of Educational Psychology*, 108(8), 1078–1092.

Kunnan, A.J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context: Proceedings of the ALTE Barcelona Conference* (pp. 27–48). Cambridge: Cambridge University Press.

Landerl, K., Ramus, F, Moll, K., Lyytinen, H., Leppanen, P. et al. (2013). Predictors of developmental dyslexia in European orthographies with varying complexity. *Journal of Child Psychology and Psychiatry*, 54, 686–694.

Lewandowski, L.J., Lovett, B.J., & Rogers, C.L. (2008). Extended time as a testing accommodation for students with reading disabilities: Does a rising tide lift all ships?. *Journal of Psychoeducational Assessment*, 26(4), 315–324.

Lewandowski, L., Cohen, J., & Lovett, B.J. (2013). Effects of extended time allotments on reading comprehension performance of college students with and without learning disabilities. *Journal of Psychoeducational Assessment*, 31(3), 326–336.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, R H., & Bates, D. (2017). Balancing Type I Error and Power in Linear Mixed Models. *Journal of Memory and Language*, 94, 305–315.

Pae, T.I. (2019). A simultaneous analysis of relations between L1 and L2 skills in reading and writing. *Reading Research Quarterly*, 54, 109–124.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, 8, 293–304.

Perfetti, C.A., Zhang, S., & Berent, I. (1992). Reading in English and Chinese: Evidence for a "Universal" Phonological Principle. In R. Frost and L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 227–248). Amsterdam: North-Holland.

R Core Team (2019). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. <http://www.R-project.org/>.

Sireci, S.G., Scarpati, S.E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, 75, 457–490.

Sokal, L., & Vermette, L. A. (2017). Double time? Examining extended testing time accommodations (ETTA) in postsecondary settings. *AHEAD Association Online*, 30, 185–200.

Sokal, L., & Wilson, A. (2017). In the nick of time: A pan-Canadian examination of extended testing time accommodation in post-secondary schools. *Canadian Journal of Disability Studies*, 6(1), 28–62.

Sparks, R.L., & Ganschow, L. (1993). The impact of native language learning problems on foreign language learning: Case study illustrations of the linguistic coding deficit hypothesis. *Modern Language Journal*, 77, 58–74.

Sparks, R., & Ganschow, L. (2001). Aptitude for learning a foreign language. *Annual Review of Applied Linguistics*, 21, 90–111.

Spenceley, L.M., Wood, W.L., Valentino, M., & Lewandowski, L.J. (2019). Predicting the extended time use of college students with disabilities. *Journal of Psychoeducational Assessment*, DOI: 0734282919848588.

Tabachnick, B.G., & Fidell, L.S. (2001). *Using multivariate analysis*. London: Pearson.

Tóth, D., Csépe, V., Vaessen, A. & Blomert, L. (2014). *3 DM – H: A diszlexia differenciáldiagnózisa. Az olvasás és helyesírás kognitív elemzése. Technikai kézikönyv [The differential diagnosis of dyslexia: The cognitive analysis of reading and spelling: Technical manual]*. Nyíregyháza: Kogentum Kft.

Tunmer, W.E., & Chapman, J.W. (2012). The simple view of reading redux. *Journal of Learning Disabilities*, 45, 453–466.

Tsagari, D., & Spanoudis, G. (Eds.) (2013). *Assessing L2 students with learning and other disabilities*. Newcastle upon Tyne, UK: Cambridge Scholars Publishing.

Vaessen, A., Bertrand, D., Tóth, D., Csépe, V., Faísca, L., Reis, A., & Blomert, L. (2010). Cognitive development of fluent word reading does not qualitatively differ between transparent and opaque orthographies. *Journal of Educational Psychology*, 102(4), 827–842.

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 reading comprehension: A componential analysis. *Journal of Educational Psychology* 96,(1),19–30.

Ziegler, J., & Goswami, U. (2005). Reading acquisition, developmental dyslexia and skilled reading across languages: a psycholinguistic grain size theory. *Psychological Bulletin*, 31, 3–29.

Zuriff, G.E. (2000). Extra examination time for students with learning disabilities: An examination of the maximum potential thesis. *Applied Measurement in Education*, 13(1), 99–117.

British Council Assessment Research Awards and Grants

If you're involved or work in research into assessment, then the British Council Assessment Research Awards and Grants might interest you.

These awards recognise achievement and innovation within the field of language assessment and form part of the British Council's extensive support of research activities across the world.

TIME-EXTENSION AND THE SECOND LANGUAGE READING PERFORMANCE OF CHILDREN WITH DIFFERENT FIRST LANGUAGE LITERACY PROFILES

AR-G/2019/1

Judit Kormos
Michael Ratajczak

ARAGs RESEARCH REPORTS ONLINE

ISSN 2057-5203