Bilingualism: Language and Cognition

cambridge.org/bil

How useful are native language tests for research with advanced second language users?

Hanke Vermeiren and Marc Brysbaert

Department of Experimental Psychology, Ghent University, Ghent, Belgium

Research Article

Cite this article: Vermeiren, H., & Brysbaert, M. (2024). How useful are native language tests for research with advanced second language users? *Bilingualism: Language and Cognition*, **27**, 204–213. https://doi.org/10.1017/S1366728923000421

Received: 29 April 2022
Revised: 17 May 2023
Accepted: 19 May 2023
First published online: 23 June 2023

Keywords:
vocabulary test; second language research; reading comprehension; general knowledge

Corresponding author:
Hanke Vermeiren, KU Leuven campus Kulak Kortrijk, Etienne Sabbelaan 51, 8500 Kortrijk, Belgium
E-mail: hanke.vermeiren@kuleuven.be

Abstract

We investigated the extent to which language tests developed for native speakers (L1) can be used with advanced speakers of a second language (L2). We compared the performance of Dutch–English bilinguals with that of native English speakers on a series of English language tests, looking at vocabulary knowledge, crystallized intelligence, reading comprehension, and reading speed. It was found that advanced L2 speakers know fewer L2 words than native speakers and take longer to read texts but perform equally well on text comprehension. Tests optimized for native English-speakers predicted text comprehension less well than tests better adapted to the skill level of the bilinguals (which include the Lextale test). An exploratory graphical analysis suggested that L2 users' performance on challenging vocabulary tests, along with performance on an English author recognition test, forms a distinct cluster – arguably also measuring interest in English language and culture besides knowledge in general (also called crystallized intelligence).

Introduction

Language researchers require valid tests, either commercially developed or created for research purposes. A recurring question with these tests is the extent to which they can be used with populations other than those for which they were developed. For example, few language tests have been developed specifically for advanced second language (L2) users of English. This raises the question to what extent language tests developed for native speakers (L1 speakers) can be used with this group?

As L2 speakers become more proficient, they become more and more like L1 speakers. This can be seen, for example, in brain imaging studies. Whereas novice L2 speakers usually show additional activation in regions related to executive functions, proficient L2 speakers exhibit activation largely confined to the language processing areas observed in L1 speakers (Cargnelutti et al., 2019; Grant et al., 2015; Indefrey, 2006).

At the same time, differences between L1 and L2 speakers are likely to remain. For a start, the vast majority of L2 speakers know fewer words in the target language than matched L1 speakers, because their exposure to the target language is less extensive. The difference in L2 vs. L1 vocabulary size is one of the largest in psychology, often not requiring statistical tests to be seen. For instance, Izura et al. (2014) developed a Spanish vocabulary test on which L1 speakers scored 53.9/60 (N = 91, SD = 6.6) and L2 undergraduates 11.9/60 after attending Spanish courses for a year at university and often for more than a year in high school (N = 123, SD = 17.9). This is a standardized effect size of d = 2.6 in favor of L1 speakers. Ferré and Brysbaert (2017) gave the test to bilinguals in Catalonia, who are among the most balanced bilinguals in the world (Guasch et al., 2011).[1] Participants who described themselves as Spanish–Catalan bilinguals (with Spanish as the dominant language) scored 53.2/60 (N = 70, SD = 5.6), whereas participants who described themselves as Catalan-Spanish bilinguals scored 48.9 (N = 86, SD = 7.1). This was still a standardized effect size of d = .66, well above the average effect size of d = .4 in psychological research (Brysbaert, 2019a).

Low-frequency words in particular are often not mastered by L2 speakers in countries where the language is not a dominant L1 language, because they are not encountered frequently enough (Cobb, 2007). In addition, L2 speakers often start learning the second

[1]Two languages are used in Catalonia: Catalan and Castilian. Both are Romance languages that originated in different kingdoms of Spain after the fall of the Roman Empire, one in the north-east and one in the center. Over the centuries, Castilian came to dominate and is now often called Spanish. In the Barcelona and Valencia region, however, Catalan remains the first language for many people, even though they are perfectly fluent in Spanish for most topics. Other people came to Catalonia from other regions in Spain and consider Spanish to be their first language.

This article has earned badges for transparent research practices: Open Data and Open Materials. For details see the Data Availability Statement.

language in school, around the end of primary school or at the beginning of secondary school, meaning that they are less likely to master L2 words referring to childhood experiences because of low exposure to these words. Brysbaert et al. (2021) reported particularly large differences between English L1 and L2 speakers in word knowledge for themes such as animals, tools, flowers, fabrics, and clergy. The differences were much smaller for words referring to distance, relatives, noisy things, science, and reading materials. Even if L2 users acquire L2 child words as they become more proficient, the semantic information associated with these concepts remains lower, as the words were learned in an educational environment rather than in everyday social interactions. There is also evidence that words acquired in early childhood continue to be processed more efficiently than later acquired words throughout life (Brysbaert & Ellis, 2016).

Another difference between L1 and L2 speakers is that L2 users benefit from cognates. Cognates are words in L1 and L2 that have the same meaning and a similar form, because they have a common origin. An example is the word 'apple' for Dutch–English bilinguals. Cognates are acquired more easily than other words by L2 learners and continue to be processed more efficiently in proficient L2 speakers (van Hell & Dijkstra, 2002). Conversely, a number of words are false friends in L1 and L2 (e.g., 'room', which also exists in Dutch but means 'cream') or have non-overlapping meanings/senses (e.g., the word 'isolatie' in Dutch is used to refer both to isolation and insulation), often leading to L2 misunderstandings. The result of the diverging relationships between L1 and L2 words is that not all L2 words are equally easy to learn.

Finally, there is evidence that L2 speakers are less sensitive to the emotional finesses associated with words in L2 (Costa et al., 2017; Hadjichristidis et al., 2019; Sulpizio et al., 2019), meaning that they may not learn the fine distinctions existing between emotion-related words (e.g., Boyd & Goldberg, 2011).

The above factors help explain why word knowledge differs between L1 and L2 speakers, even for advanced L2 speakers who can express themselves well in L2. This raises the question of how well language tests developed for native speakers can be used for advanced L2 speakers. We investigate this question by giving a set of language tests developed for native English speakers to a group of Dutch–English bilingual university students. How large will the differences be and, more importantly, how do the correlations of the tests compare between L2 and L1 speakers?

The tests were compiled by Vermeiren et al. (2022), who wanted to validate a free English vocabulary test for university undergraduates, so that it could be used in psycholinguistic research. The goal turned out to be more challenging than foreseen and after five studies Vermeiren et al. (2022) ended up with three vocabulary tests of 50 multiple choice items each, together with four reading comprehension tests, an author recognition test, and a general knowledge test. These are interesting tests to administer to English L2 speakers.

The reasons why Vermeiren et al. (2022) had to run five studies were twofold. First, for some time it seemed as if there were two types of words, which the authors provisionally termed: (1) unfamiliar words for specialized information (StuVoc1), and (2) unfamiliar words for familiar experiences (StuVoc2). The first type of words consisted of academic words related to advanced knowledge (e.g., polynomial); the second type primarily consisted of infrequent synonyms of familiar words (e.g., baneful instead of harmful). In the end, however, both groups of words correlated highly with each other and the two vocabulary tests did not load on separate factors.

The second reason why Vermeiren et al. (2022) ended up with three vocabulary tests was that their first attempt to build a test with unfamiliar words for familiar experiences proved to be too easy for their sample (university students). As a result, the test did not differentiate well among the participants, most of whom achieved high scores. The authors speculated that the test (called StuVoc3) might be a good test for less proficient groups, such as younger participants and advanced L2 speakers, but did not verify this claim.

In addition to the vocabulary tests, Vermeiren et al. (2022) used four reading comprehension tests as a validation criterion, because vocabulary size correlates well with reading comprehension (e.g., Calloway et al., 2022). Four tests were used because the individual tests did not score highly on reliability (coefficients of r = .5 – .6). These tests will be described in more detail in the method section.

The author recognition test was included in Vermeiren et al. (2022) to have an objective measure of language exposure. Scores on this test are known to correlate with vocabulary knowledge, also in L2 users (Kim & Krashen, 1998; McCarron & Kuperman, 2022; Moore & Gordon, 2015). Participants were asked to indicate which fiction authors they knew.

The general knowledge test was developed as a separate measurement of crystallized intelligence. Vocabulary tests are often included in intelligence tests to measure crystallized intelligence (cultural knowledge stored in long-term memory) along with general knowledge questions. Care was taken that the general knowledge questions did not rely heavily on vocabulary knowledge. Thus, the questions were not of the type "What do you call a horse-like animal with black stripes?" but of the type "Why is it warmer in the summer than in the winter?"

Using structural equation modelling, Vermeiren et al. (2022) observed that their tests loaded on two correlated latent factors. The first was crystallized intelligence, with significant loadings of the vocabulary tests, the general knowledge test, and (surprisingly) the author recognition test. The second factor was reading comprehension, with significant loadings of the four reading comprehension tests. The crystallized intelligence factor correlated r = .6 with the reading comprehension factor. A third, largely independent, factor was formed by the reading rates in the comprehension tests.

Figure 1 gives a different summary of the findings obtained by Vermeiren et al. (2022, Study 5). Instead of a factor analysis, it shows a network analysis.[2] The analysis is based on an algorithm developed by Golino and Epskamp (2017), which includes a cluster analysis (Walktrap community detection) and an exploratory graph analysis. Such an analysis is equivalent to an exploratory factor analysis, but does not assume the existence of latent factors. As a result, it provides a more theory-neutral picture of the relationships between the various measurements. In order to deal with issues such as data skewness and the presence of outliers in a uniform and principled way, the network is based on Spearman correlations instead of Pearson correlations (Bishara & Hittner, 2015; de Winter et al., 2016; Isvoranu & Epskamp, 2021).

Christensen and Golino (2021) added a bootstrapping option to the exploratory graph analysis. This option makes it possible to examine how stable the obtained solution is, by generating

[2]Vermeiren et al. (2022) also included the Nelson-Denny vocabulary test, which was not included in the present study because of the costs involved. Therefore it has been omitted from Figure 1. Vocabulary test StuVoc3 was not part of Study 5 of Vermeiren et al. So, it is not included either.

https://doi.org/10.1017/S1366728923000421 Published online by Cambridge University Press

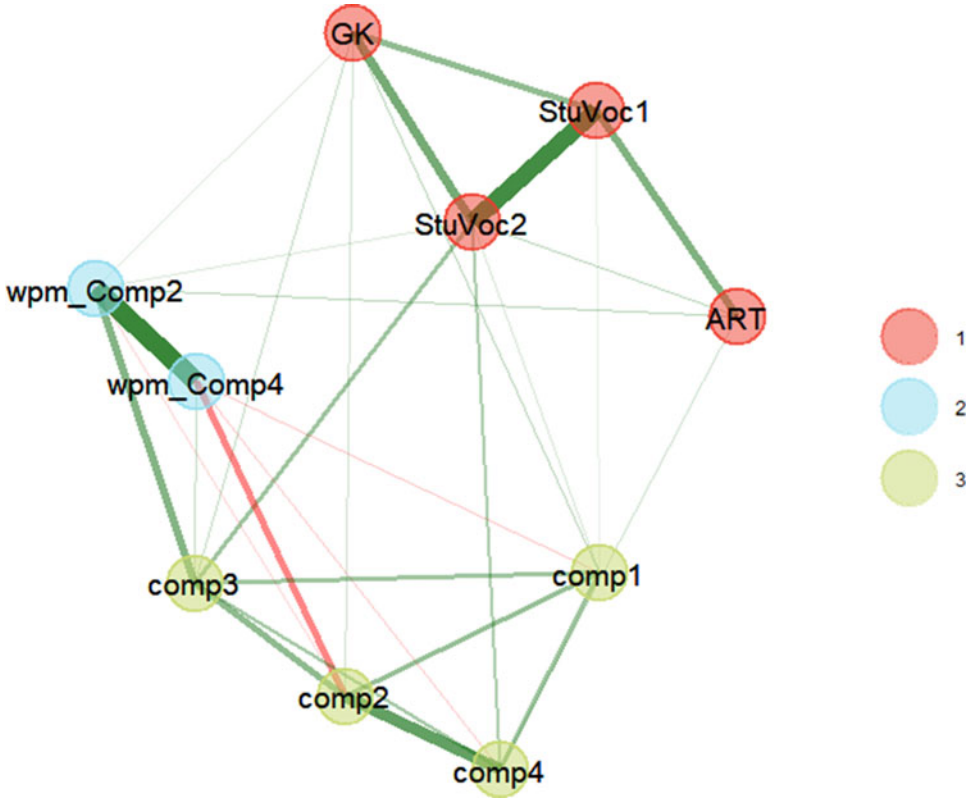

Figure 1. Dominant exploratory graph analysis of Vermeiren et al. (2022, Study 5) based on the R package EGAnet (Christensen & Golino, 2021). This analysis shows that the variables group in three clusters: (1) crystallized intelligence including the two vocabulary test (StuVoc1, StuVoc2), the author recognition test (ART), and the general knowledge test (GK); (2) reading rates of reading comprehension test 2 and 4 (wpm_Com2, wpm_Com4); and (3) a cluster formed by the accuracy scores on the four reading comprehension tests administered (comp1-comp4). The thickness of the lines illustrates the partial correlations between the variables after conditioning on the other variables, which is equivalent to the predictive quality between two nodes that would be obtained in multiple regression.

new stimulus sets based on the one obtained and repeating the analysis on these sets. The parametric bootstrapping showed that the distinction between the cluster related to crystallized intelligence and the cluster related to reading comprehension was present in all simulations. The distinction between reading comprehension and reading rate was present in 88% of the 500 simulations. In the remaining 12%, reading rates formed a single cluster with reading comprehension.

The data of Vermeiren et al. (2022) are an interesting test bed to compare performance of advanced L2 speakers to that of L1 speakers. Will they show the same network? How much will performance differ for the various tests? For instance, it could be hypothesized that performance on StuVoc1 will be rather good in L2 relative to L1, because many of the words are learned in academia. The same may not be true for StuVoc2, as this includes more infrequent synonyms of familiar words.

Altogether, our study was an exploratory study, to see how Dutch–English bilinguals would perform on the various tests developed for native English speakers by Vermeiren et al. (2022), and how their network of correlations would compare to that depicted in Figure 1. The main stimulus materials were the same as in Vermeiren et al. (2022, Studies 3 and 5). All tests are available for use at https://osf.io/2xyzn/?view_only=43fcaf9053404ee195041791ef08d013.

Method

Participants

Participants were first-year psychology students at Ghent University. These students typically have B2 and C1 levels in the CEFR framework developed by the Council of Europe (described in https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions). For reading, these levels mean that the participants can understand most general texts and even specialized texts not related to their field of interest. In addition, they can appreciate distinctions in style. They do, however, not attain the highest level (which is C2). Typically, their comprehension of not too complex English texts is at the same level as that of L1 speakers, but their reading rate is slower (Dirix et al., 2020; Kuperman et al., 2022). This research group is interesting for the current study, because the participants are among the most advanced English L2 speakers tested in language research, but still perform substantially worse than native speakers. They typically started formal education in English as a second language at the age of 13, although many already knew quite some English vocabulary before that age, resulting from out-of-school contextual learning (De Wilde et al., 2020).

To make sure that we had rather stable data, we tested some 200 participants, so that the 95% confidence intervals around the correlation estimates were smaller than r = .15 (discussed in more detail by Vermeiren et al., 2022, who used the same sizes in their studies). Responses were gathered from 210 participants, of whom 5 had to be excluded due to data suggesting low effort in the last part of the study (fast responses, repeated answer alternatives). Participant exclusion was decided before data analysis began, in order not to affect the findings.[3] Students who completed the whole study received three course credits (equivalent to three hours testing). Since we sampled from psychology students, the gender distribution was skewed (female: 89%, male: 11%).

[3]There may be a case to exclude an additional 14 participants outside the range of 17-21 years used in Vermeiren et al. (2022). Doing so does not alter the conclusions, but makes the EGA network less stable. These participants were included to make maximal use of the available evidence.

https://doi.org/10.1017/S1366728923000421 Published online by Cambridge University Press

Materials

In addition to the tests included in Vermeiren et al. (2022, Study 5), we decided to include two more tests. The first was Lextale (Lemhöfer & Broersma, 2012). This test is often used in psycholinguistic research to assess the proficiency of the participants. It consists of words and non-words, and participants have to indicate which words they know. They are penalized for yes-answers to non-words. Lemhöfer and Broersma (2012) reported high correlations with placement tests, but more recent publications have raised doubt about the word recognition format used. Based on a meta-analysis, Zhang and Zhang (2020) concluded that L2 recall tests correlate more with reading comprehension than L2 recognition tests, and that tests about the meaning of the words correlate more with reading comprehension than tests about the word forms. Given that the word/non-word format of Lextale is a recognition test of word forms, Zhang and Zhang's (2020) review raises doubts about the usefulness of the test to predict reading comprehension, even though the format was not included in the meta-analysis.

Evidence in line with Zhang and Zhang (2020) was published by McLean et al. (2020). These authors predicted English language comprehension in Japanese university students using four different test formats: form recognition (yes/no checklists), form recall (translating L1 words into L2), meaning recognition (recognizing the correct L2 option in a multiple-choice test), and meaning recall (translating L2 words into L1). Although form recognition had the best reliability for short tests, it correlated less well with text comprehension. The highest correlation was r = .67 (for tests with more than 100 items). In line with Zhang and Zhang (2020), McLean et al. (2020) reported that the best predictor of reading comprehension was meaning recall (r = .78), followed by form recall (r = .75), and meaning recognition (r = 0.71). Given the findings of Zhang and Zhang (2020) and McLean et al. (2020), it was interesting to include Lextale in our study, so that we could collect more information about this vocabulary test.

Finally, we included a Dutch (L1) vocabulary test based on meaning recognition (L1 multiple-choice format) to see how this test would fit within the network, given that an L1 vocabulary test is often used as the best measure of crystallized intelligence (Schipolowski et al., 2014).

Vocabulary tests StuVoc1, StuVoc2, StuVoc3. Participants were presented with the three vocabulary tests developed by Vermeiren et al. (2022). Each test consisted of 50 visually presented target words with four response alternatives. Participants had to select the right alternative. StuVoc1 is a vocabulary test that correlated particularly well with general knowledge, the other measure of crystallized intelligence (as can be seen in Figure 1). It also correlated highly with the author recognition test, although this was not expected. StuVoc2 contains words acquired before the age of 14 but not by everyone, because the concepts can easily be described by other words. The last test, StuVoc3, was compiled on the same principles as StuVoc2 (i.e., mostly consisting of early acquired words), but turned out to be too easy for native English-speaking university students, resulting in a ceiling effect. It will be interesting to see whether this test works better with L2 speakers, as speculated by Vermeiren et al. (2022).

Author recognition test (ART3)

The author recognition test was initially developed by Stanovich and West (1989) as an objective measure of reading frequency. The idea was that people who read a lot know the names of more fiction authors than people who do not read much. We used the third version of the original ART test (Vermeiren et al., 2022), including more recent, popular youth authors, hence ART3. The ART3 includes 60 fiction author names and 30 non-author names. The names are presented in random order and participants have to indicate which authors they know. Surprisingly, the test correlated more with measures of crystallized intelligence than with measures of reading comprehension (as shown in Figure 1), suggesting that participants are more likely to know the authors' names on the basis of their general cultural knowledge, than because they have effectively read the books of all the authors.

General knowledge (GK)

In Vermeiren et al. (2022), a test with 80 quiz questions was composed, covering a large number of topics. Questions were visually presented multiple-choice items with four response options. After the first administration of the test, the stimuli were pruned to the 65 best ones (based on an item response theory analysis).

Reading comprehension tests

Four comprehension tests were used by Vermeiren et al. (2022, Study 5). All texts were typical for expository texts seen in newspaper articles, Wikipedia entries, and introductory textbooks. That is, they did not contain many unexplained low-frequency words and the syntax was rather simple.

Comp1 was a 1056-word explanatory text about the ice ages designed by Griffin et al. (2008) and followed by 24 true/false statements. A reliability of .48 was found.

Comp2 was a comprehension test created by Vermeiren et al. (2022). It consisted of 14 short expository texts (100-150 words) covering a wide range of topics, each followed by 3 questions with four response alternatives. Reliability was .65.

Comp3 was published by Kane and Miyake (2007). It consisted of 20 short text fragments (40–114 words). Participants had to indicate which of five response alternatives was the most likely continuation of the fragment. This test had a time limit of 10 minutes, which was too short to finish all questions. Kane and Miyake (2007) reported a reliability of alpha = .76. Vermeiren et al. (2022) found a reliability of .82 (ICC). However, the test had significant additional loading on crystallized intelligence and reading speed (the faster participants read, the higher their scores), likely because of the time pressure in the test.

The last comprehension test, *Comp4*, was compiled by Yeari et al. (2015) and contained 10 rather short texts (314 to 458 words), followed by five true or false statements. The texts covered a variety of topics ranging from reality TV to Einstein. Vermeiren et al. (2022) reported a reliability of .60.

Reading rate

In addition to accuracy, comp1, comp2 and comp4 also gave estimates of reading rate, expressed as words per minute. Reading rate could not be calculated for Comp3, because participants had to answer questions while reading the text (and did not complete all texts).

Lextale

In addition to the test used by Vermeiren et al. (2022), we also included the Test for Advanced Learners of English (Lextale) (Lemhöfer & Broersma, 2012). This vocabulary test is often used in psycholinguistic research with participants speaking

https://doi.org/10.1017/S1366728923000421 Published online by Cambridge University Press

English as L2. In the Lextale test participants are presented with 40 words and 20 nonwords and have to indicate whether they know the English word or not (for specific instructions, see Lemhöfer & Broersma, 2012 or https://lextale.com/takethetest.html). Performance was calculated as percentage yes-responses to words minus percentage yes-responses to nonwords. Typical performance for students at Ghent University is 70-75%. Vermeiren et al. (2022) presented the test in Study 3 to L1 speakers, which resulted in average performance of 90% with rather small SD (see Table 1 below).

Dutch vocabulary test

Finally, we also included a Dutch L1 meaning recognition test. Vander Beken et al. (2018) developed such a test with 75 multiple choice items, each having a visually presented target word with four response alternatives. A reliability of .84 (Cronbach's alpha) was reported and the test correlated .6 with English L2 proficiency. After an item analysis, the test was shortened to 40 questions without loss of information.

All tests are available for use at https://osf.io/2xyzn/files/osfstorage

Results

Raw data and analysis code can be found at https://osf.io/2xyzn/files/osfstorage. First, we present the descriptive statistics of the various tests. We give means, standard deviations, and reliability coefficients. Reliability is based on the intraclass coefficient ICC2k of the Psych R package (Revelle, 2021). This gives the reliability under the assumption that the items are a random sample from a homogeneous population. In general, it is slightly lower than Cronbach's alpha, which in turn is slightly lower than McDonald's omega. As such, the number is a lower estimate of reliability (differences with omega mostly smaller than .05). We also compared the performance of the L2 participants tested in the present study to that of the L1 participants in Vermeiren et al. (2022).

Descriptive statistics vocabulary tests

Table 1 shows the descriptive statistics for all the vocabulary tests we ran. For the StuVoc and Lextale tests, they also show the performance of the L1 speakers in Vermeiren et al. (2022). As expected, the difference between L1 and L2 speakers was smallest for the StuVoc1 test, which included many words mostly encountered in academic texts. The difference was larger for the other vocabulary tests, in particular for StuVoc3, where the standard deviation of the L1 speakers was small due to a ceiling effect. The differences show that the present participants performed at a level clearly below that of native speakers (as indicated above, Izura et al., 2014, found a difference of d = .7 between highly proficient Spanish L2 speakers and L1 speakers, and a difference of d = 2.6 between beginning learners and native speakers). All tests had good reliability (ICC > .75).

Descriptive statistics author recognition test and general knowledge test

Table 2 shows the results of ART3 and GK. Again, a comparison is made with the findings of the L1 participants in Vermeiren et al. (2022, Study 5). As can be seen, the L2 speakers performed slightly worse than the L1 speakers. For the author recognition test, this is understandable as the authors in the test were specifically selected for English-speaking test takers. The lower performance for the general knowledge test was less expected. It could mean that the present sample of undergraduate students was slightly less knowledgeable than the Prolific sample of Vermeiren et al. (2022). Prolific is an online site with quality control to recruit participants for research, where characteristics can be defined that participants must meet in order to participate (e.g., having English as a first language and being of a certain age; Peer et al., 2022). It could also mean that fact retrieval is a bit more difficult in L2 than in L1.

Descriptive statistics comprehension tests

To make sure that the texts were read for comprehension, trials with reading rates above 668 words per minute (wpm) were omitted. Such reading rates are 2.5 times higher than the average reading rate of 240 wpm and are typical for text scanning rather than text reading (Brysbaert, 2019b). No lower limit was set, as L2 speakers are known to have slower reading rates (Dirix et al., 2020; Kuperman et al., 2022). All in all, 302 of the 22,156 reading comprehension observations (1.36%) had to be dropped. Omitted texts are a particular problem for Comp1, as this test includes only one text. Six participants had excessively fast reading speeds. Because Comp3 had a fixed completion time of 10 min and induced time pressure, no reading rate could be calculated for this test. To deal with missing data due to the removal of answers outside of the reading speed range, mean scores instead of sum scores were calculated for comp2 and comp4.

Table 3 lists the findings and compares them to the L1 speakers of Vermeiren et al. (2022). Calculations were based on the sum scores (Comp1 and Comp3) or mean scores in the case of missing data (Comp2 and Comp4). Intraclass correlation based on mixed effects modeling is used to deal with missing data when calculating reliability. As can be seen in Table 3, reading comprehension was similar for L1 and L2 readers. Only for Comp3, a lower score was found for L2 readers, probably due to the time limit applied in this test. A less interesting aspect was that the reliability of two tests (Comp1 and Comp3) was lower in the present study than in Vermeiren et al. (2022).

Descriptive statistics reading rate

Table 4 gives the descriptive statistics of reading rate in words per minute. On average participants read 147 words per minute, which is considerably lower than the range of the 220–260 words per minute observed in L1 readers (Brysbaert, 2019b).

Correlation analysis

Because missing data as a result of too high reading rates affected both reading accuracy and reading rate, which are correlated, reading rate of Comp1 was dropped and the six missing observations for reading comprehension were imputed with de default options of the R package mice version 3.14.0 (seed = 500; van Buuren, 2021), as was done in Vermeiren et al. (2022, Study 5). Mice stands for multiple imputation by chained equations. This analysis allows researchers to impute missing values based on the observed values for a given individual and the relations observed in the data for the other participants (Azur et al., 2011; Schafer & Graham, 2002). Table 5 shows the results of the correlational analysis (Spearman correlations).

https://doi.org/10.1017/S1366728923000421 Published online by Cambridge University Press

Table 1. Descriptive statistics and reliability measures for the vocabulary tests and comparisons with the results of L1 speakers reported in Vermeiren et al. (2022). N = 205 for the present study, N = 182 for the StuVoc tests in Vermeiren et al., and N = 196 for the Lextale test. Effect size is Cohen's d with 95% confidence interval.

	L2 (current study)			L1 (Vermeiren et al., 2022)			Difference
Test	M	sd	ICC	M	sd	ICC	standardized
StuVoc1 (max = 50)	24.9	7.00	0.77	31.7	8.61	0.86	−0.88 [−1.09, −0.67]
StuVoc2 (max = 50)	21.7	7.44	0.80	34.0	9.29	0.89	−1.47 [−1.70, −1.25]
StuVoc3 (max = 50)	31.0	7.95	0.84	44.9	4.09	0.79	−2.16 [−2.41, −1.91]
Lextale (max = 100)	74.9	11.16	0.77	90.5	7.00	0.69	−1.66 [−1.88, −1.43]
Dutch Voc (max = 40)	22.6	6.74	0.81				

Table 2. Descriptive statistics and reliability measures for the author recognition test and the general knowledge test (N = 205). Comparison with the results of the L1 speakers reported in Vermeiren et al. (2022, Study 5; N = 182). Effect size is Cohen's d with 95% confidence interval.

	L2 (current study)			L1 (Vermeiren et al., 2022)			Difference
Test	M	sd	ICC	M	sd	ICC	standardized
ART3 (max = 100)	53.4	8.52	0.85	57.3	9.36	0.87	−0.46 [−.66, −.26]
GK (max = 65)	38.7	6.76	0.69	42.0	7.75	0.78	−0.46 [−.66, −.25]

Table 3. Descriptive statistics and reliability measures for the reading comprehension tests (N = 205). Comparison with the results of the L1 speakers reported in Vermeiren et al. (2022, Study 5; N = 182). Effect size is Cohen's d with 95% confidence interval.

	L2 (current study)			L1 (Vermeiren et al., 2022)			Difference
Test	M	sd	ICC	M	sd	ICC	standardized
Comp1 (max = 24)	16.9	2.53	0.29	17.3	2.81	0.44	−0.15 [−.35, +.05]
Comp2 (max = 1.00)	0.72	0.13	0.74	0.74	0.12	0.70	−0.16 [−.36, +.04]
Comp3 (max = 20)	5.2	3.03	0.64	9.0	4.66	0.82	−0.98 [−1.19,−.77]
Comp4 (max = 1.00)	0.71	0.10	0.59	0.69	0.10	0.58	+0.20 [−.00, +.40]

Table 4. Descriptive statistics and reliability measures for the reading rates (N = 205). Comparison with the results of the L1 speakers reported in Vermeiren et al. (2022, Study 5; N = 182). Texts with reading rates above 668 words per minute were omitted. Effect size is Cohen's d with 95% confidence interval.

	L2 (current study)			L1 (Vermeiren et al., 2022)			Difference
Test	M	sd	ICC	M	sd	ICC	standardized
Comp1	159	83.7	NA	261	122.2	NA	−0.98 [−1.20, −0.77]
Comp2	145	54.1	0.92	224	63.3	0.89	−1.35 [−1.57, −1.13]
Comp4	149	49.7	0.89	266	94.9	0.86	−1.57 [−1.80, −1.34]

Several aspects are noteworthy in Table 5. First, the correlation between StuVoc1 and StuVoc2 (ρ = .70) is lower than the one found with native English speakers (r = .85), in line with the observation that L2 speakers performed relatively better on StuVoc1 than on StuVoc2. Second, of the three StuVoc tests, StuVoc3 correlated most with reading comprehension. The correlations with StuVoc1 and StuVoc2 are much lower than those observed in L1 speakers, and would have forced us to conclude that the relationship between word knowledge and reading comprehension is much lower for advanced L2 speakers than L1 speakers, if these had been the only tests at our disposal. Third, Lextale did not correlate less with reading comprehension than the StuVoc tests. This goes against the concerns raised against the Lextale format by Zhang and Zhang (2020) and McLean et al. (2020). Unlike the other vocabulary tests, Lextale correlated positively with reading rate: Participants with high scores on Lextale tended to read faster than participants with low scores. Fourth, the general knowledge test correlated well

https://doi.org/10.1017/S1366728923000421 Published online by Cambridge University Press

Table 5. Spearman correlations between the various variables (N = 205). Correlations of .19 and more are printed in bold because they are significant at .01.

Test	1	2	3	4	5	6	7	8	9	10	11	12
1. StuVoc1	–											
2. StuVoc2	**.70**	–										
3. StuVoc3	**.49**	**.54**	–									
4. LexTale	**.31**	**.45**	**.61**	–								
5. VocNL	**.39**	**.43**	**.48**	**.31**	–							
6. ART	**.26**	**.36**	**.19**	**.19**	**.22**	–						
7. GK	**.35**	**.39**	**.54**	**.52**	**.41**	**.19**	–					
8. Comp1	.12	.15	**.34**	**.26**	**.26**	.01	**.31**	–				
9. Comp2	**.27**	**.30**	**.55**	**.45**	**.38**	**.21**	**.45**	**.55**	–			
10. Comp3	.12	**.21**	**.33**	**.41**	**.25**	.08	**.26**	**.24**	**.34**	–		
11. Comp4	**.23**	**.19**	**.51**	**.40**	**.30**	.09	**.40**	**.48**	**.63**	**.32**	–	
12. Wpm_comp2	.03	.07	.09	**.21**	−.06	.01	.06	−.01	−.10	**.31**	−.08	–
13. Wpm_comp4	.06	.03	.00	.17	.02	−.08	−.01	−.07	−.17	**.30**	−.11	**.67**

with the vocabulary tests and with the reading comprehension tests, in line with the observation that reading comprehension is helped by having background knowledge of the topic. Fifth, as in Vermeiren et al. (2022) the author recognition test correlated more with tests of crystallized intelligence than with reading comprehension tests. Most of the time, the correlations between the author recognition test and the reading comprehension tests are not significant. Sixth, scores on the reading comprehension tests correlated with each other, despite some low reliabilities. Comp3 again was the odd one out, due to its significant positive correlation with reading rate. Students who read fast were at an advantage for this test, whereas they tended to be at a disadvantage in the other comprehension tests. Finally, the L1 vocabulary test correlated positively with all other variables, except for reading rate. On average, participants who did well on Dutch vocabulary also did well on English tests. This is in line with the assumption that all test scores rely on (crystallized) intelligence.

Network analysis

As in Figure 1, we used the exploratory graph analysis developed by Golino and Epskamp (2017) and Christensen and Golino (2021), to get a better picture of the relations between the different test scores. The algorithms were again run on the Spearman correlation matrix. Figure 2 shows the outcome based on 500 iterations of the parametric bootstrapping algorithm (a very similar analysis was obtained, if only the dataset itself was analyzed).

The outcome of the analysis is largely compatible with the L1 network from Figure 1, except for one change. A new cluster emerged, including the two demanding English vocabulary tests and the English author recognition test. The most likely origin of this cluster is differences in interest in English literature and culture (and corresponding differences in motivation to learn more about them). The cluster is related to a cluster of tests measuring crystallized intelligence but, surprisingly, very little with performance on the English reading comprehension tests we used.

The bootstrap analysis indicated that the distinct cluster of reading speed was found in virtually all 500 analyses. The distinct cluster of advanced English knowledge was found in 80% of the analyses; in the remaining the tests joined the crystallized intelligence cluster. A separate reading comprehension cluster was obtained in 83% of the analyses; in the remaining 17% the tests joined the crystallized intelligence cluster.

As for the individual tests, the most unsteady test was Comp3. In 47% of the simulations it joined the reading comprehension cluster, in 20% the crystallized intelligence cluster, in 31% the reading speed cluster, and in 2% of the simulations it formed a cluster on its own. Remember that Comp3 was the reading test in which participants had to choose as many best paragraph continuations as possible in 10 minute's time (Kane & Miyake, 2007). Vermeiren et al. (2022) already observed that this test is not a pure measurement of reading comprehension.

Discussion

The present study was an exploratory study, to examine how advanced English L2 speakers perform relative to English L1 speakers on a battery of English tests, involving vocabulary, reading comprehension, reading rate, and general knowledge. The L2 participants were first-year psychology students at a Belgian university, where students are expected to have a good command of English (e.g., many courses make use of English textbooks). The L1 participants were native speakers of similar age, tested online via Prolific.

The main findings are in line with a robust pattern described in the literature. The L2 participants obtain Lextale scores between 65 and 85% (Table 1), expected of B2 and C1 speakers (Lemhöfer & Broersma, 2012) and typically obtained with first-year Dutch-speaking undergraduates. This is well below the level of native speakers, with effect sizes that do not require sophisticated statistical analysis to be observed (Table 1). As with other non-native students, English proficiency increases throughout the years of post-secondary education (McCarron & Kuperman, 2022), in particular when study materials are in English. The students are capable of studying in English, but need more time to do so (Table 3; see also Dirix et al., 2020; Kuperman et al., 2022) and perform worse in recall exams, such as essay writing (de Vos et al., 2020; Dhaene & Woumans,

https://doi.org/10.1017/S1366728923000421 Published online by Cambridge University Press

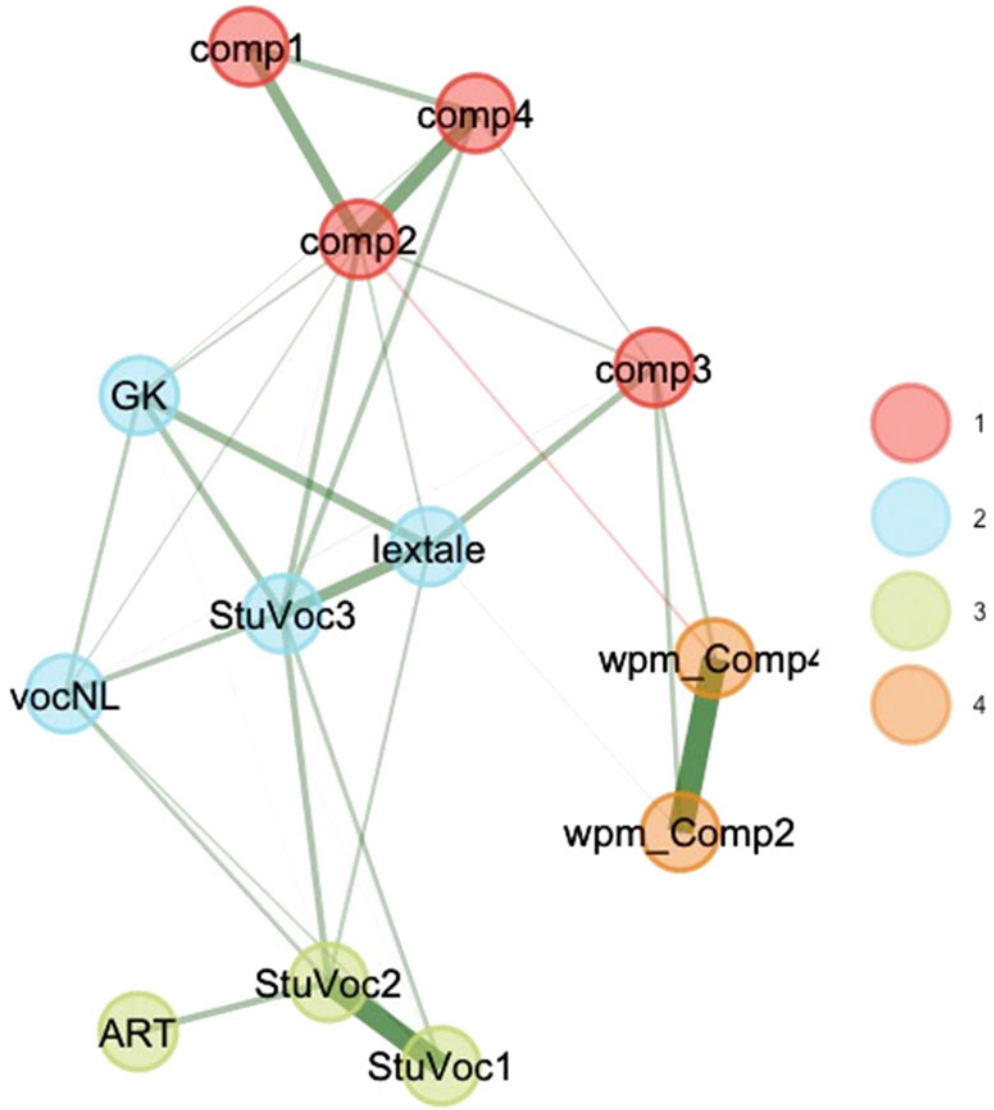

Figure 2. Result from EGA analysis based on parametric bootstrapping (500 iterations; seed = 500). The tests formed four clusters: (1) reading comprehension including the four reading comprehension tests (comp1-comp4), (2) crystallized intelligence including general knowledge (GK), Lextale, StuVoc3, and the Dutch vocabulary test (vocNL), (3) reading speed including the reading rates of comp2 and comp4 (wpm_Comp), and (4) an English knowledge cluster, including the advanced vocabulary tests StuVoc1 and StuVoc2, and the author recognition test (ART). Lines between nodes indicate partial correlations after conditioning on the other variables, which is equivalent to the predictive quality between two nodes that would be obtained in multiple regression. See the online article for a version with colour.

2023; Vander Beken & Brysbaert, 2018). They perform better with recognition tests, where native-like accuracy performance can be observed for materials that do not contain many unexplained low-frequency words (Table 4; see also Vander Beken et al., 2018, 2020).

Some students know more English than others. In particular, they perform well on vocabulary tests that are challenging for L1 speakers too (StuVoc1 and StuVoc2). Interestingly, this did not lead to better reading comprehension for the materials and the untimed recognition tests we used, over and above the contribution of crystallized intelligence. Indeed, the two challenging vocabulary tests formed a separate cluster, along with the English author recognition test. This is a reminder that crystallized intelligence (measuring cultural knowledge) forms a hierarchy of several areas of interest, with individual differences in the importance attached to each (Steger et al., 2019). Some people are more interested in mastering the English language and culture than others. It can be expected that these individuals will perform better than their L2 peers in more challenging test situations than those tested here (e.g., limited reading time, recall) or on texts discussing topics related to English language and culture. A further observation is that performance on StuVoc1 and StuVoc2 was well correlated. This is in line with Vermeiren et al.'s (2022) observation that both tests measure the same skill.

In contrast to the advanced English vocabulary tests, a vocabulary test that better matched the participants' general mastery level (StuVoc3) correlated more with other measures of crystallized intelligence (L1 vocabulary test, general knowledge test) and comprehension of introductory non-fiction texts. The same was true for the Lextale test. The latter predicted reading comprehension as well as StuVoc3, contrary to the concerns raised against the format by McLean et al. (2020; see also Zhang & Zhang, 2020). One possibility could be that the lower predictive validity of the Lextale format is specific to Asian samples. Indeed, Lemhöfer and

https://doi.org/10.1017/S1366728923000421 Published online by Cambridge University Press

Broersma (2012) also reported lower validity indices for a Korean sample than for a Dutch sample.

A final observation of interest is that for L2 readers too reading rate is largely independent of reading accuracy. This observation has been made many times in L1 research (reviewed in Brysbaert, 2019b; see also Figure 1). Now we see the same pattern in L2 (see also Kuperman et al., 2022). This forms a challenge for further research. On the one hand, we see little correlation between reading rate and reading comprehension. On the other hand, we observe that L2 readers have slower reading rates than L1 readers but similar comprehension levels. So, they seem to have some metacognition about the relationship between reading speed and text difficulty, but they do not use this knowledge to maximize their reading comprehension. One explanation may be that readers differ in the degree of text memory they aspire to and adapt their reading rate accordingly. Another reason for the low correlation may be that fast and slow reading have at least two, conflicting origins. Indeed, slow reading can be the outcome of low skill but also of a desire to memorize and organize the text well. Similarly, fast reading can be due to both high skills and a lack of motivation to do well. It will be interesting to see whether such influences can be disentangled by experimental design.

All in all, our study in combination with Vermeiren et al. (2022) gives us a bird's eye view of the commonalities and differences between L1 and L2 readers. Returning to the question raised in the title of the article, to what extent L1 tests can be used with advanced L2 speakers, we have to conclude that our data raise concerns. Even though advanced L2 readers differ in their performance on challenging tests, these differences seem to be less related to everyday L2 expository text understanding than tests matched to the participants' proficiency level (see in particular the low correlations of StuVoc1, StuVoc2 and ART with reading comprehension). This is important information to keep in mind when one wants to use a vocabulary test to gauge L2 language proficiency.

Acknowledgements. All research was conducted at Ghent University. The corresponding author is currently affiliated with Faculty of Psychology and Educational Sciences, and Imec research group Itec, KU Leuven, Leuven, Belgium

Data availability. The data that support the findings of this study are openly available in OSF at https://osf.io/2xyzn/files/osfstorage

References

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. *International Journal of Methods in Psychiatric Research*, *20*(1), 1–62.

Bishara, A. J., & Hittner, J. B. (2015). Reducing bias and error in the correlation coefficient due to nonnormality. *Educational and Psychological Measurement*, *75*(5), 785–804.

Boyd, J. K., & Goldberg, A. E. (2011). Learning what not to say: The role of statistical preemption and categorization in a-adjective production. *Language*, *87*(1), 55–83. https://doi.org/10.1353/lan.2011.0012

Brysbaert, M. (2019a). How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables. *Journal of Cognition*, *2*(1), 16.

Brysbaert, M. (2019b). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language. 109*.

Brysbaert, M., & Ellis, A. W. (2016). Aphasia and age-of-acquisition: Are early-learned words more resilient? *Aphasiology*, *30*, 1240–1263.

Brysbaert, M., Keuleers, E., & Mandera, P. (2021). Which words do English non-native speakers know? New supernational levels based on yes/no decision. *Second Language Research*, *37*(2), 207–231.

Calloway, R., Helder, A., & Perfetti, C. (2022). A measure of individual differences in readers' approaches to text and its relation to reading experience and reading comprehension. *Behavior Research Methods*, *55*(2), 899–931.

Cargnelutti, E., Tomasino, B., & Fabbro, F. (2019). Language brain representation in bilinguals with different age of appropriation and proficiency of the second language: A meta-analysis of functional imaging studies. *Frontiers in Human Neuroscience*, *13*, 154.

Christensen, A. P., & Golino, H. (2021). Estimating the Stability of Psychological Dimensions via Bootstrap Exploratory Graph Analysis: A Monte Carlo Simulation and Tutorial. *Psych*, *3*, 479–500.

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology 11*(3), 38–63.

Costa, A., Vives, M. L., & Corey, J. D. (2017). On language processing shaping decision making. *Current Directions in Psychological Science*, *26*(2), 146–151.

de Vos, J. F., Schriefers, H., & Lemhöfer, K. (2020). Does study language (Dutch versus English) influence study success of Dutch and German students in the Netherlands?. *Dutch Journal of Applied Linguistics*, *9*(1–2), 60–78.

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2020). Learning English through out-of-school exposure. *Which levels of language proficiency are attained and which types of input are important? Bilingualism: Language & Cognition*, *23*, 171–185.

de Winter, J. C., Gosling, S. D., & Potter, J. (2016). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods*, *21*(3), 273–290.

Dhaene, S., & Woumans, E. (2023). Text recall and use of advance organisers in first and second language. *Studies in Second Language Acquisition*, *45*(1), 264–275.

Dirix, N., Vander Beken, H., De Bruyne, E., Brysbaert, M., & Duyck, W. (2020). Reading Text When Studying in a Second Language: An Eye-Tracking Study. *Reading Research Quarterly*, *55*(3), 371–397.

Ferré, P., & Brysbaert, M. (2017). Can Lextale-Esp discriminate between groups of highly proficient Catalan-Spanish bilinguals with different language dominances? *Behavior Research Methods*, *49*, 717–723.

Golino, H., & Epskamp, S. (2017). Exploratory Graph Analysis: A new approach for estimating the number of dimensions in psychological research. *PLoS ONE*, *12*.

Grant, A. M., Fang, S. Y., & Li, P. (2015). Second language lexical development and cognitive control: A longitudinal fMRI study. *Brain and Language*, *144*, 35–47.

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self- explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition*, *36*, 93–103.

Guasch, M., Sanchez-Casas, R., Ferre, P., & García-Albea, J. E. (2011). Effects of the degree of meaning similarity on cross-language semantic priming in highly proficient bilinguals. *Journal of Cognitive Psychology*, *23*(8), 942–961.

Hadjichristidis, C., Geipel, J., & Keysar, B. (2019). The influence of native language in shaping judgment and choice. *Progress in Brain Research*, *247*, 253–272.

Indefrey, P. (2006). A meta-analysis of hemodynamic studies on first and second language processing: Which suggested differences can we trust and what do they mean?. *Language Learning*, *56*, 279–304.

Isvoranu, A.-M., & Epskamp, S. (2021). Which estimation method to choose in network psychometrics? Deriving guidelines for applied researchers. *Psychological Methods*.

Izura, C., Cuetos, F., & Brysbaert, M. (2014). Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicológica*, *35*, 49–66.

Kane, M. J., & Miyake, T. M. (2007). The validity of "conceptual span" as a measure of working memory capacity. *Memory & Cognition 35*, 1136–1150.

Kim, H., & Krashen, S. (1998). The author recognition and magazine recognition tests, and free voluntary reading as predictors of vocabulary development in English as a foreign language for Korean high school students. *System*, *26*(4), 515–523.

https://doi.org/10.1017/S1366728923000421 Published online by Cambridge University Press

Kuperman, V., Siegelman, N., Schroeder, S., Acartürk, C., Alexeeva, S., Amenta, S., Bertram, R., Bonandrini, R., Brysbaert, M., Chernova, D., Da Fonseca, S. M., Dirix, N., Duyck, W., Fella, A., Frost, R., Gattei, C. A., Kalaitzi, A., Lõo, K., Marelli, M… & Usal, K. A. (2022). Text reading in English as a second language: Evidence from the Multilingual Eye-Movements Corpus. *Studies in Second Language Acquisition*, 1–35.

Lemhöfer, K., & Broersma, M. (2012). Introducing Lextale: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*. *44*, 325–343.

McCarron, S. P., & Kuperman, V. (2022). Is the author recognition test a useful metric for native and non-native English speakers? An item response theory analysis. *Behavior Research Methods*, *53*(5), 2226–2237.

McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*.

Moore, M., & Gordon, P. C. (2015). Reading ability and print exposure: Item response theory analysis of the author recognition test. *Behavior Research Methods*, *47*(4), 1095–1109.

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*(4), 1643–1662.

Revelle, W. (2021). psych: Procedures for Personality and Psychological Research, R package Version 2.1.9. Retrieved from https://CRAN.R-project.org/package=psych

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*, 147–177.

Schipolowski, S., Wilhelm, O., & Schroeders, U. (2014). On the nature of crystallized intelligence: The relationship between verbal ability and factual knowledge. *Intelligence*, *46*, 156–168.

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, *24*(4), 402–433.

Steger, D., Schroeders, U., & Wilhelm, O. (2019). On the dimensionality of crystallized intelligence: A smartphone-based assessment. *Intelligence*, *72*, 76–85.

Sulpizio, S., Toti, M., Del Maschio, N., Costa, A., Fedeli, D., Job, R., & Abutalebi, J. (2019). Are you really cursing? Neural processing of taboo words in native and foreign language. *Brain and Language*, *194*, 84–92.

van Buuren, S. (2021). Package 'mice' Version 3.14.0. Cran: https://cran.r-project.org/web/packages/mice/mice.pdf

Vander Beken, H., & Brysbaert, M. (2018). Studying texts in a second language: The importance of test type. *Bilingualism: Language and Cognition*, *21*(5), 1062–1074.

Vander Beken, H., Woumans, E., & Brysbaert, M. (2018). Studying texts in a second language: No disadvantage in long-term recognition memory. *Bilingualism: Language and Cognition*, *21*(4), 826–838.

Vander Beken, H., De Bruyne, E., & Brysbaert, M. (2020). Studying texts in a non-native language: A further investigation of factors involved in the L2 recall cost. *Quarterly Journal of Experimental Psychology*, *73*(6), 891–907.

van Hell, J. G., & Dijkstra, T. (2002). Foreign language knowledge can influence native language performance in exclusively native contexts. *Psychonomic Bulletin & Review*, *9*(4), 780–789.

Vermeiren, H., Vandendaele, A., & Brysbaert, M. (2022). Validated tests for language research with university students whose native language is English: Tests of vocabulary, general knowledge, author recognition, and reading comprehension. *Behavior Research Methods*, *55*(3), 1036–1068. https://doi.org/10.3758/s13428-022-01856-x.

Yeari, M., van den Broek, P., & Oudega, M. (2015). Processing and memory of central versus peripheral information as a function of reading goals: Evidence from eye-movements. *Reading and Writing: An Interdisciplinary Journal*, *28*(8), 1071–1097.

Zhang, S., & Zhang, X. (2020). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, https://doi.org/10.1177/1362168820913998

https://doi.org/10.1017/S1366728923000421 Published online by Cambridge University Press