Cognition and Neurosciences

Lexize: A test to quickly assess vocabulary knowledge in Finnish

ROSA SALMELA,[1] MINNA LEHTONEN,[2,3] STEFANO GARUSI[1] and RAYMOND BERTRAM[2,4]

[1]*Department of Psychology, Åbo Akademi University, Turku, Finland*
[2]*Department of Psychology and Speech-Language Pathology, University of Turku, Turku, Finland*
[3]*MultiLing Center for Multilingualism in Society across the Lifespan, University of Oslo, Oslo, Norway*
[4]*Linguistic Anthropology Laboratory, Tomsk State University, Tomsk, Russia*

Salmela, R., Lehtonen, M., Garusi, S. & Bertram, R. (2021). Lexize: A test to quickly assess vocabulary knowledge in Finnish. *Scandinavian Journal of Psychology.*

Proficiency in a language is strongly related to how well and how many words one knows. Vocabulary knowledge correlates with reading comprehension and general communication ability. Due to the increasing amount of research within the field of psycholinguistics and second language acquisition in Finnish, a standardized test to objectively measure Finnish vocabulary knowledge is called for. Lexize is such a test. It was modeled after LexTALE (Lemhöfer & Broersma, Behaviour Research Methods, 44:325–343, 2012), which was developed to measure vocabulary knowledge of English as a second language using visual lexical decision (VLD). Lexize is a VLD-based online test for Finnish that consists of 102 items. By comparing performance of L1 and L2 speakers of Finnish, Lexize was validated, returning considerable differences between test scores in native and non-native speakers. For non-native speakers there was a large range of test scores, correlating strongly with exposure to Finnish and self-ratings. In native speakers, test scores correlated with self-ratings, Finnish school grades, and age. In this group, higher Lexize scores were associated with a higher education level. We conclude that Lexize is a useful tool to assess Finnish vocabulary knowledge for non-native speakers and to some extent for native speakers. Lexize is available for free use at https://psyk.abo.fi/LexizeWeb/#/.

Key words: L2, Finnish, lexicon, language proficiency, lexical decision, vocabulary knowledge, vocabulary test.

Rosa Salmela, Department of Psychology, Åbo Akademi University, Fabriksgatan 2, Åbo 20500, Finland. E-mail: rosa.salmela@abo.fi

INTRODUCTION

Proficiency in a language is strongly related to vocabulary knowledge. Knowing a lot of words entails better reading comprehension and general communication ability (Laufer & Ravenhorst-Kalovski, 2010; Nation, 1993; Staehr, 2008). Vocabulary development correlates strongly with phonological and syntactic development (Bates & Goodman, 1999; Gathercole & Baddeley, 1989). In L2 speakers, a clear relationship between vocabulary knowledge and more general language abilities has been established. For instance, Hilton (2008) found a strong correlation between L2 speech fluency and L2 vocabulary knowledge. A systematic review of Jeon and Yamashita (2014) depicts the strong correlation between L2 vocabulary knowledge and L2 reading comprehension. This is underlined by the more recent systematic review of Zhang and Zhang (2020), who also established the strong relationship between L2 vocabulary knowledge and listening comprehension.

With respect to vocabulary knowledge, often a distinction is made between vocabulary breadth and vocabulary depth. Breadth of vocabulary refers to the number of words known, and depth of vocabulary refers to the richness of word knowledge (Schmitt, 2014). This distinction resonates with lexical organization accounts like the Lexical Quality Hypothesis (Perfetti & Hart, 2001) which holds that words are represented as integrated patterns of orthographic, phonological, and semantic representations. The orthographic and phonological specification are less context-dependent than the semantic specification, which needs more time to emerge and to be refined. Tran, Tremblay and Binder (2020) argue that the form-related representations can be thought of as the breadth dimension, while the semantic part would convey depth of vocabulary knowledge.

Typically, both dimensions of vocabulary knowledge correlate with each other and both dimensions correlate with reading comprehension (Li & Kirby, 2014). Li and Kirby found evidence that vocabulary breadth is a better predictor of reading comprehension, while vocabulary depth better predicts ability to write summaries, a measure of deeper text processing. Tran *et al*. (2020) found that for low literacy readers vocabulary breadth explains 67% of the variance in a reading comprehension task.

In sum, vocabulary breadth is strongly associated with both oral and written language proficiency. A quick, validated test that assesses vocabulary breadth and that can be used as a proxy for language proficiency is therefore of great value for language researchers, especially in the field of bilingualism and second language (L2) acquisition. Moreover, a validated vocabulary knowledge test could be utilized in L2 instruction and in clinical work as well. Several vocabulary knowledge tests already exist in English and many other languages, but not in Finnish. Due to the increasing number of studies in psycholinguistics and second language acquisition in Finnish, as well as an increasing amount of immigration, a standardized test to objectively measure Finnish vocabulary knowledge, and thus proficiency, is called for. The current study intends to accomplish that.

Vocabulary knowledge tests in English

There are a number of standardized vocabulary knowledge tests in English. Many of them measure receptive vocabulary knowledge

Section Editor: Mikael Heimann

at the level of vocabulary breadth by means of recognition. This is typically tested by the visual lexical decision paradigm (VLD). In this paradigm, the participants are presented with letter strings one by one on a computer screen. Some of the letter strings are genuine words in the target language, for example, *cat* in English, and some of them are pseudowords, for example, *kilp*. For each trial, the participant must indicate whether the target string is an existing word or not. The decision is typically made by a button-press, that is, "yes" for a real word and "no" for a pseudoword.

The advantage of this type of testing lies in the quick and easy administration, which is often important in experimental settings. Sometimes recognition tests have been criticized for lack of ecological validity, as they tend to yield higher scores than recall tests, which may suggest that people may be prone to guessing the correct answer (Gyllstad, Vilkaite & Schmitt, 2015; McLean, Kramer & Stewart, 2015). However, according to the lexical quality hypothesis (Perfetti & Hart, 2001), these higher scores may just reflect partial word knowledge being assessed, as words can be known at the formal level without having much semantic specification. In the following, we will introduce the most widely used vocabulary tests in English, and then present Lexize, a test to measure vocabulary knowledge in Finnish.

Vocabulary Levels Test

Perhaps the most widely used vocabulary test in English is the Vocabulary Levels Test (VLT). This test was originally developed by Paul Nation in the 1980s (Nation, 1983), and subsequently revised by Schmitt, Schmitt and Clapham (2001). The test is not computerized and uses a type of form-meaning matching where the learner has to select a correct definition for a given word. Four levels of difficulty are created by using word frequency as a proxy of difficulty. The four frequency levels are the 2,000, 3,000, 5,000 and 10,000 level. These levels reflect the frequency level a word belongs to, for example, words that belong to the 2,000 level are among the 2,000 most frequently used words in English. The final section of the VLT tests academic vocabulary knowledge and is not frequency-based (Schmitt *et al*., 2001). At each level, the tasks are presented in clusters of six words, and the learner has to match three of them with the three definitions provided. Each level consists of ten clusters. The test estimates the participants' language proficiency on the basis of the number of correctly identified items at the different difficulty levels. The VLT thus taps into the initial stages of learning form-meaning links, more at the level of vocabulary breadth than vocabulary depth. More precisely, it establishes at which frequency level a language user recognizes the majority of the words (Schmitt, 2010). The test has been a useful tool for placing students into ability groups in language, but there are only a few published studies that have investigated the validity of the instrument (Read, 1988; Schmitt *et al*., 2001). Schmitt *et al*. (2001) found that the items in VLT distinguished well between better and weaker learners. Read (1988) found that knowledge of words in lower word frequency levels was associated with higher language proficiency level. However, according to some studies, the possibility of blind guessing of the correct responses cannot be ruled out (e.g., Webb, 2008).

Eurocentres Vocabulary Size Test

The Eurocentres Vocabulary Size Test (EVST) by Meara and Jones (1987) is another well-known test for measuring vocabulary size in English. The test uses the VLD task to assess lexical knowledge and, similarly to VLT, it utilizes word frequency as an indicator of item difficulty.

The EVST is divided into 10 blocks each corresponding to a frequency band of 1,000 words. The test starts of at the highest frequency band and if the participant scores highly, s/he will be tested on the next band. The test continues until performance drops below a given threshold. The test samples 10 genuine words and 10 pseudowords at each level and takes approximately 10 min to run. The total test score is calculated by summing up the scores for each block of words (Meara & Jones, 1987). In contrast to VLT, the test is fully computerized and generates scores for the participants automatically. The scoring system is based on Signal Detection Theory models (Zimmerman *et al*., 1977) and takes into account both the hit rate of the genuine words and the false alarm rates of the incorrectly answered pseudowords. Regardless of the advantages of easy and quick administration, some concerns have been expressed. It has been noted for instance that the EVST does not consistently correlate with the other widely used test, the VLT (see e.g., Cameron, 2002; Mochida & Harrington, 2006). More specifically, it has been found that the EVST is particularly good at discriminating at higher levels of proficiency, whereas the VLT is better at the lower end but loses some discriminatory power at the higher end of proficiency (Meara & Miralpeix, 2016).

Vocabulary Size Test

The Vocabulary Size Test (VST; Nation & Beglar, 2007) is a 140-item test, designed to measure written receptive knowledge of the first 14,000 words of English. Multiplying the test score with 100 provides an estimation of the actual vocabulary size, that is, a score of 100 would thus imply a vocabulary size of 10,000. The test is suitable for both L1 and L2 learners. Testees are asked to select the best definition from four choices of target words presented in sentence context. The VST thus measures knowledge of the form-meaning connection, and to a smaller degree concept knowledge. Word richness is not assessed, so the test pertains to the level of vocabulary breadth (Nation & Beglar, 2007).

Lexical Test for Advanced Learners of English

The Lexical Test for Advanced Learners of English (LexTALE, Lemhöfer & Broersma, 2012) is a validated vocabulary test for measuring English vocabulary proficiency. The test served as a model for subsequent versions in other languages (LexTALE_FR, Brysbaert, 2013; Lextale-Esp, Izura, Cuetos & Brysbaert, 2014; LEXTALE_CH, Chan & Chang, 2018; LexITA, Amenta, Badan & Brysbaert, 2020) and was therefore chosen as basis for Lexize.

The original LexTALE test (Lemhöfer & Broersma, 2012) was designed to provide researchers with a practical and objective measure of proficiency by assessing testees' receptive lexical knowledge. It is intended for researchers studying participants with an advanced level of English as a second language in an experimental setting. Similarly to EVST and VLT, LexTALE utilizes word frequency to divide the test words into different difficulty levels, which, in turn, are used to assess vocabulary knowledge. The test is based on a simple unspeeded VLD task, consists of 40 English words and 20 pseudowords and can be completed within 5–10 min. The primary outcome measure is the percentage of correct responses over both words and pseudowords, corrected for the unequal proportion of words and pseudowords in the test. The test was validated with external criteria, that is, translation tasks, proficiency tests, and experimental data in VLD tasks. The LexTALE scores correlated well with participant performance in these experimental paradigms (Lemhöfer & Broersma, 2012).

Subsequently, parallel tests were designed and validated for French and Spanish. The French test, LexTALE_FR (Brysbaert, 2013), is an identical unspeeded VLD task consisting of 92 items. The validation of the test began with a larger set of items ($n = 120$) from varying frequency bands, after which item assessment was conducted via Classical Test Theory (CTT) and Item Response Theory (IRT). Both methods are widely used in psychometrics, as they estimate the internal consistency of a given test. This is done by assessing how individual test items relate to general performance. CTT estimates item discrimination power based on the notion that the test items usually correlate with the total test score. IRT is considered to offer advantages over CTT, as the item statistics in IRT are independent of the groups from which they are estimated. This means also that the scores describing the testee's vocabulary knowledge is not tied to the test difficulty and the test items may be matched to ability levels where they function best (Hambleton & Jones, 1993). Thus, items whose identification accuracy correlated poorly with the participants' ability level were excluded from the final test. The external criterion for test validity was established by comparing LexTALE_FR scores to self-ratings and years of education in French.

The Spanish version of the test was developed (Lextale-Esp) by Izura *et al.* (2014) and is almost identical by design to LexTALE_FR. It originally consisted of 180 items with 90 pseudowords and 90 words from six frequency bands; after item assessment, the final set consisted of 90 items including 60 words and 30 pseudowords. Lextale-Esp discriminated well at the high and the low end of Spanish proficiency and returned a large difference between the vocabulary size of Spanish native and non-native speakers.

Similar results are reported for Mandarin Chinese (LEXTALE_CH, Chan & Chang, 2018) and for a vocabulary test derived from LexTALE in Italian (LexITA; Amenta *et al.*, 2020). After item assessment, both of these tests also consist of 90 items including 60 words and 30 pseudowords.

Vocabulary tests in Finnish

In experimental settings, time- and resource-effective test administration is often a priority. However, standardized vocabulary knowledge tests to assess Finnish language proficiency of immigrants or adult native (L1) speakers have not been developed yet. For native L1 speakers, the level of language proficiency has been assessed mainly by self-ratings, but their validity has been shown to be relatively poor (e.g., Delgado *et al.*, 1999). With regard to L2 speakers, measures like exposure and age of acquisition are often used as a proxy of language proficiency, but objective proficiency level measurement still relies heavily on extensive and often time-consuming language tests. Thus, there is a clear need in the field of language research for standardized, quick tests that can give insight into participants' language proficiency in Finnish.

There are a few studies in Finnish that investigated lexical development of mono- and bilingual children by non-standardized methods. For example, Honko (2013) studied vocabulary development of second-generation immigrants during the first 6 years of elementary school (age 7–12), on the basis of frequency analyses in self-made word recognition tasks and analyses of written narratives. Niiranen (2008) studied Norwegian-Finnish bilingual school children (age 12–15) who acquired Finnish and Norwegian simultaneously since childhood. In this study, Niiranen used a modified version of the verb identification task of EVST (Meara & Jones, 1987; Meara, 1996). Saarela (1997), in turn, studied lexical development of L1 Finnish elementary school children (ages 8–14) living in Finland by assessing the maturity of their vocabulary in school essays. The maturity was defined by three variables: abstractness, diversity, and nuances. Due to the variety of methods used, comparison of the results may be challenging.

In psycholinguistic studies on L2 Finnish, language proficiency is often assessed via self-ratings or questionnaires assessing exposure to L2 (e.g., Lehtonen & Laine, 2003; Portin, Lehtonen & Laine, 2007). Sometimes university language centers' tailored course placement tests are used in studies as well (e.g., Kimppa, Shtyrov, Hut, Hedlund, Leminen & Leminen, 2019). The only standardized test package for assessing L2 Finnish, *Kielo* (Tani, 2008), is widely used in integration programs, which include Finnish courses that are offered to unemployed immigrants in Finland. However, Kielo is not often used in language research, as it is an extensive assessment tool. Kielo consists of four separate sections covering all language modalities: writing, speaking, listening and reading. The test takes several hours to complete and has to be marked manually. This requires a significant number of hours for both the participant and the experimenter and makes the test suboptimal for experimental settings.

In sum, a variety of methods have been used to assess lexical knowledge and language proficiency in Finnish, which makes it hard to compare the results. Moreover, the most extensive assessment tool is typically too time-consuming for research purposes. This status quo underlines the need for a test that can be used across studies and that provides an adequate and quick estimation of L2 Finnish proficiency. The standardized vocabulary knowledge test in Finnish presented in the current study aims to realize precisely that.

Lexize: a test to quickly assess vocabulary knowledge in Finnish

In the present study, our objective is to create a quick and standardized vocabulary knowledge test for experimental studies in Finnish that can be used as a proxy for general language proficiency in Finnish. This vocabulary test is modeled after the LexTALE test by Lemhöfer and Broersma (2012). We opted for a new name that refers to vocabulary size, as Lexize scores are thought to capture lexical knowledge with lower or higher scores typically corresponding with smaller or larger vocabulary size, and being, therefore, reflective of general language proficiency (Hilton, 2008; Jeon & Yamashita, 2014; Zhang & Zhang, 2020). Lexize was designed to measure vocabulary knowledge from low-proficient to high-proficient speakers of Finnish. The main objective was to design a tool that could distinguish between different levels of lexical knowledge of L2 speakers. However, as Finnish proficiency may also widely vary among L1 speakers, for instance as a function of age or educational level, we also tested whether Lexize can be used to detect different levels of lexical knowledge among native Finnish speakers. As the test is linked to a comprehensive questionnaire, an additional aim is to use the test to examine underlying factors explaining one's vocabulary knowledge in different population groups.

As in the previous LexTALE studies (Brysbaert, 2013; Izura *et al*., 2014), the current Lexize test went through a careful item selection and item analysis procedure. We started off with a large number of stimuli pretested with a group of L1 and a group of L2 speakers and retained only the stimuli with substantial discriminatory power, that is, the stimuli that had a good ability to distinguish between participants of different levels of lexical knowledge. The discriminatory power was assessed via CTT and IRT analysis (see section "Lexical Test for Advanced Learners of English"). After this we analyzed the L1 and L2 test scores on the basis of the retained items. In order for the Lexize test to be a useful, reliable, and valid vocabulary knowledge test, it would need to return considerable differences between L1 and L2 performance. In addition, it would need to differentiate within groups, in both L1 and L2 speakers. Hence, we validated the test by assessing the difference between L1 and L2 speakers, as well as the correlations between the test score and proficiency self-ratings (L1 and L2 speakers), education level, and age (L2 speakers), and school grade for the Finnish language (L1 speakers). These variables were chosen as external criteria because several studies suggest they correlate strongly with language competence and lexical knowledge (Bowers & Vasilyeva, 2011; Brysbaert, Stevens, Mandera & Keuleers, 2016; Grøver, Lawrence & Rydland, 2018; Herschensohn, 2009; Keuleers, Stevens, Mandera & Brysbaert, 2015; Park, Lautenschlager, Hedden, Davidson, Smith & Smith, 2002; Rydland, Grøver & Lawrence, 2014).

In what follows, we will describe the development and validation of the Lexize test in detail. First, we will describe the initial version of the test and then provide details on how the item assessment was conducted via the CTT analysis and the IRT analysis. These analyses will be presented in the Methods section. In the Results section, we will present the correlations between Lexize scores and the above-mentioned variables for both L1 and L2 speakers.

METHOD

In this section, we first describe the participants and the participant selection procedure. Then we describe the questionnaire that every participant filled in prior to the Lexize vocabulary test. Subsequently, we describe the exact administration of Lexize followed by a description of the lexical items and item selection criteria. Finally, the analyses that led to the final item selection will be presented in detail.

Participants

Altogether 309 participants performed the test, but after screening for missing data and reported language disorders, 276 participants (117 L1 speakers and 159 L2 speakers) were selected for further analyses. The participant characteristics for the L1 and L2 groups are specified in Table 1. L1 participants were recruited from the University of Turku and vocational schools in Turku area, Finland. L2 participants were recruited from the Swedish speaking Åbo Akademi University student email-lists and L2 adult education centers in Turku area. The language background of the L2 speakers was heterogeneous, consisting of 60 different languages. The largest language groups had Swedish (n = 68) or Russian (n = 28) as their L1.

Questionnaire

Before stimulus presentation, participants were subjected to a questionnaire including a number of background questions as well as several ratings that were used for cross-validation and further analyses. More specifically, we asked a number of questions pertaining to gender, age, grade for Finnish in the final secondary school year (L1 speakers), Finnish proficiency ratings, exposure to Finnish, possible language

Table 1. *Lexize participant characteristics*

	L1 (n = 117)			L2 (n = 159)		
	M	*SD*	Mdn	*M*	*SD*	Mdn
Age	22.5	6	21	29.2	8.1	27
Exposure to Finnish[a]	22.5	6	21	11.8	10.8	7
Self-rating for Finnish[b]	8.4	1.2	9	5.4	2.2	5
School grade for Finnish[c]	8.4	1.1	9	NA	NA	NA
Gender	Male: 44, Female: 65, Other: 8			Male: 37, Female: 120, Other: 2		
Native language	Finnish			Several (n = 60)		
Education level[d]	Primary = 35, Secondary = 46, University = 36			Primary = 3, Secondary = 34, University = 122		

[a] Years lived in Finland.
[b] Self-rating based on scale 1–10.
[c] Final assessment in Finnish in the end of comprehensive school (national scale 4–10).
[d] Primary: highest completed level is comprehensive school; Secondary: highest completed level is high school or vocational school; University: highest completed level is university level education (bachelor or master).

deficits, profession, and educational level. With respect to proficiency, participants were asked to rate their Finnish language proficiency on a scale from 1 (very weak) to 10 (outstanding). With respect to exposure, participants were asked to indicate how many years they had lived in Finland. The questions are listed in Appendix S2.

Administration

Lexize is implemented as a web-based vocabulary knowledge test and is available at https://psyk.abo.fi/LexizeWeb/#/.[1] The participants were provided with the web address and they performed the test on a mobile device or PC. The type of device should not affect the results, as reaction times were not relevant in this study. The participants took the test either as a part of ongoing L2 research at the University of Turku, or, in case they were recruited from the email-lists, at home or other remote location. L1 participants from vocational institutions took the test locally in a school class, as part of a Finnish course. The procedure began with a background questionnaire, after which the actual vocabulary test, a VLD task, took place. The VLD task was preceded by instructions, in which it was explained that letter strings would appear on the screen one by one and that the participant had to decide for each letter string whether it was a word or not by pressing the yes- or no-button. Each letter string was preceded by a fixation point that was on the screen for 500 ms. At the end of the test the participant was informed about the number of correct and incorrect answers and a provisionary level of expertise was indicated (e.g., novice, or advanced). There was no time limit for the individual items or for the task as a whole. There were 88 words and 44 pseudowords in the test. Progress in the task could be followed by a timeline. Although some participants took the test at a remote location, cheating was considered unlikely as the participants could get an estimation of their own language skill in an objective manner and performance was not linked to any specific reward. However, in these cases, the possibility that some participants used external resources cannot be completely ruled out.

Upon invitation, the participants were explained the purpose and procedure of the experiment, as well as their rights as a participant. They were also informed that by clicking the link they will give consent for participation in the experiment. The initial experimental webpage repeated the purpose of the experiment and the questionnaire and once more pointed out that data will be processed anonymously. At the end of the experiment the participants had the choice to click a button with “send results” which transmitted the data to a password-protected network drive of the server at Åbo Akademi University. They were informed they could refrain from sending the data without further explanation in case they did not want to share it with the researchers.

The Lexize vocabulary test

The initial version of Lexize consisted of 132 items, 88 of which were Finnish words and 44 phonotactically legal Finnish pseudowords. The real words and the lexical statistics were retrieved from a Finnish newspaper corpus comprising 22.7 million word forms by using the lexical search program WordMill (Laine & Virtanen, 1999). Words were selected from six different frequency bands. The selection included 10 words with frequency <1 per million (pm); 24 words with frequency 1–5 pm; 22 words with frequency 5–10 pm; 18 words with frequency 10–20 pm; 10 words with 20–100 pm; and 4 words with frequency more than 100 pm. This division in frequency bands and the selection procedure follows the procedure used in other LexTALE studies. The majority of words were nouns ($n = 64$), and the rest of the words were verbs ($n = 9$) and adjectives ($n = 15$). All the selected words were monomorphemic in order to avoid that knowledge of morphological structure would come to aid in word recognition. Word length ranged from 4 to 9 letters with an average of 5.7 letters.

Subsequently, a list of 44 pseudowords was compiled. To this end, we first chose 44 words that corresponded in part of speech, length and frequency to the 88 words that were selected as word stimuli for the pretest. Next, 1–3 letters in these words were changed in such a way that phonotactically legal pseudowords were created. The average bigram frequency[2] of the selected pseudowords ($M = 5.85$, $SD = 2.53$) was comparable to that of the selected words ($M = 6.19$, $SD = 2.65$), ensuring that the letter combinations of the pseudowords resemble those of the words, as shown by an independent samples t-test ($t(128) = 0.53$, $p = 0.6$). All lexical characteristics of the initial item set are listed in Appendix S1. The list of words and pseudowords was submitted to a random permutation, and this randomized list was subsequently presented to all the participants.

Item assessment

As mentioned earlier, the initial test consisted of 132 items (88 real words and 44 pseudowords). We opted for a relatively large initial set of items as it was to be expected that some items would drop out due to being unsuitable and/or having little discriminatory power. To assess whether the items were of good quality, we tested their fit by utilizing both CTT and IRT (see section “Lexical Test for Advanced Learners of English”). First, similar to Brysbaert (2013) and Izura *et al.* (2014), we calculated the point-biserial correlations in order to detect correlation between the Lexize Score of the respondents and their answer for each item separately. We used a cut-off index of 0.2, as advised by Crocker and Algina (1986), and removed the 11 words and two pseudowords with the lowest correlations. The words that correlated poorly with the overall score were *hedelmä*, *aamu*, *huone*, *talvi*, *lippu*, *ystävä*, *tumma*, *selkä*, *sairas*, *kohtelias* and *levätä* (see Appendix S1 for translations). The reason why these words correlated poorly was probably due to their high lemma frequency (>29 per million), that is, all words were from the highest or one but highest frequency band, which meant that almost all participants knew them irrespective of their Finnish proficiency level. The pseudowords that correlated poorly with the overall scores were *mahna* and *uopi*. The bigram frequency of these items was close to average (4.4 and 3.9, for *mahna* and *uopi* respectively), but for some reason most participants – even non-proficient ones – identified them correctly as pseudowords.

Next, an IRT analysis was performed for the remaining 119 items in order to get more detailed information of the difficulty and discrimination index of each item. Generally, IRT analyses can address three parameters: item difficulty (i.e., how difficult it is to achieve a 0.5 probability of a correct response for a specific item given the respondent’s overall performance in the test), discrimination (i.e., the ability of an item to differentiate among respondents given their overall performance in the test), and guessing (the probability that a respondent without any knowledge provides correct answers by chance). In the one parameter logistic model (1PL), only item difficulty is included, the two-parameter model (2PL) includes both item difficulty and discrimination, and the three-parameter model (3PL) includes all three (for more detailed description, see e.g., Mair, 2018; Raykov & Marcoulides, 2010). In contrast to Brysbaert (2013) and Izura *et al.* (2014), who conducted the analyses by using 2PL, we used a 3PL model, as it had a better fit than the 1PL and 2PL models in pairwise comparisons with ANOVA ($ps < 0.001$, see Table 2). The IRT analyses were conducted by the statistical software R (R Core Team, 2017) with package ltm (Rizopoulos, 2006).

In the 3PL IRT analysis, all remaining items turned out to be highly discriminative (Baker, 2001; $M = 3.8$, $SD = 2.9$; $M = 3.3$, $SD = 2.2$; for words and pseudowords respectively).[3] The lowest discrimination index (0.72) was detected for the pseudoword *könnä*. This index is considered

Table 2. *Likelihood ratio tables for IRT model comparison*

	AIC	BIC	log.Lik	LRT	*df*	*p* value
1PL	26,504.77	26,952.78	−13,132.4			
2PL	25,916.31	26,804.85	−12,720.2	824.46	118	<0.001
2PL	25,916.31	26,804.85	−12,720.2			
3PL	25,474.04	26,806.84	−12,380	680.28	119	<0.001

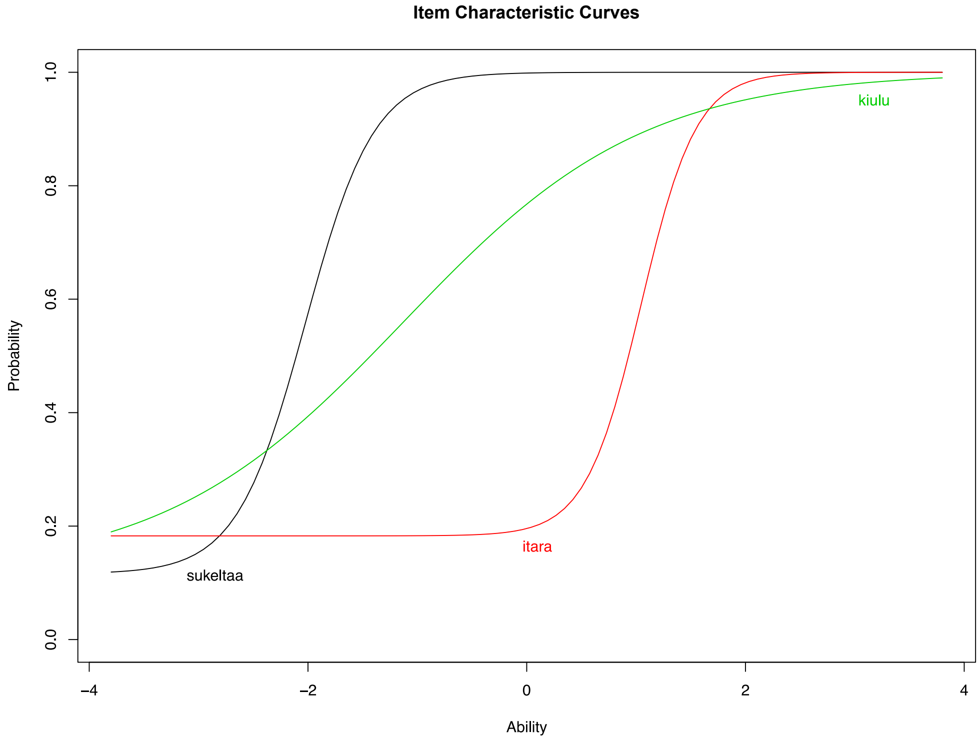

Fig. 1. Examples of item characteristic curves of three items in Lexize. Probability reflects the likelihood that a testee with a certain ability level will give a correct answer to the item. Ability is defined by the overall performance in the test. The figure shows that sukeltaa ("to dive") discriminates well in the low ability end, whereas itara ("niggard") discriminates well in the medium-high ability end. The steepness of the curve represents the item's discriminatory power, i.e., kiulu ("a pail") does not discriminate as well between ability levels as the two other items do.

moderate according to Baker (2001). Thus, no items were removed on the basis of their discriminability power. See Fig. 1 for examples.

When looking at the difficulty of the items, many of them turned out to be relatively easy ($M = -1.1$, $SD = 0.8$; $M = -0.7$, $SD = 1.1$; for words and pseudowords respectively; Baker, 2001).[4] The low difficulty of the items was detected also in the test information plot, which showed that most of the observations were centered around −1 (Fig. 2). In practice, this means that throughout the proficiency levels, the test gave most information on a relatively low level of ability (i.e., L2 speakers). This makes sense as many (but not all) of the L1 speakers' scores were close to ceiling level and in general there was less variation among them than among L2 speakers.

In order to make the test more balanced, we excluded the 17 easiest items. These included nine words (*raita*, *ateria*, *noutaa*, *kerho*, *myrsky*, *nahka*, *mänty*, *jono* and *varvas*; see Appendix S1 for translations) and eight pseudowords (*nampota*, *milsu*, *irne*, *relo*, *könnä, noimi*, *munsu* and *sukkole*). The cut-off level of exclusion (−2.04 and −2.19; for words and pseudowords respectively) was chosen, so that the final test set would consist of approximately 100 words.

With regard to the real words, the excluded items were again relatively high in lemma frequency ($M = 21$, $SD = 8.8$, range 11–35), all belonging to the one-but-highest or two-but-highest frequency band (see Fig. 3 for the distribution of frequency in words). In addition, they were concrete and imageable nouns (e.g., "stripe," "meal") that are typically taught early on in L2 instruction, making them easy for even the low level L2 speakers. With regard to the pseudowords, the bigram values did not fully explain why they were so often recognized correctly as pseudowords. However, most of them did not have a particularly high bigram value (range 2.31–6.58), which may have made them easier to recognize correctly as pseudowords. See Fig. 4 for the distribution of the bigram values in pseudowords.

When looking at the third parameter, guessing, it turned out that most of the pseudowords in the test had a high guessing index ($M = 0.5$, $SD = 0.1$), whereas the real words' guessing index was significantly below chance level ($M = 0.2$, $SD = 0.2$). According to Waller (1989), the determination of an exact value below which responses are to be omitted is constrained only by the requirement that the cut-off value must be less than or equal to the chance level of the test. We decided not to use guessing as an exclusion criterion, as this would have meant that almost all pseudowords should be excluded. The same phenomenon was observed by Brysbaert (2013) for the high difficulty index of pseudowords in the 2PL model. However, it makes sense that pseudowords have a high guessing index, especially in the L2 group as they have more lexical uncertainty in general due to smaller vocabularies and less established lexical representations. In our case, the high guessing index for the pseudowords does not have to be a problem, as the Lexize score accounts for guessing behavior by penalizing wrong answers with minus points (see section "Scoring the test"). As described afore, our motivation to test the items with the 3PL model was that it explained more variance than the 1PL or 2PL model and thus gave us more accurate information about the items' performance with regard to their difficulty and discrimination rates.

After exclusion based on the 3PL IRT analysis, we ended up with a selection of 102 items (68 words and 34 pseudowords). The discrimination power of the final items was excellent ($M = 3.9$, $SD = 2.8$), although the difficulty rate was moderately low ($M = -0.7$, $SD = 0.7$). This reflects that many items were still relatively easy and therefore are most useful in discriminating participants with lower abilities, that is, L2 speakers. Table 3 lists the lexical characteristics of the 68 words and 34 pseudowords that were selected to the final version of the Lexize vocabulary test.

Scoring the test

In line with recommendations made by Lemhöfer and Broersma (2012), Brysbaert (2013), and Izura *et al.* (2014), the test score was defined as follows:

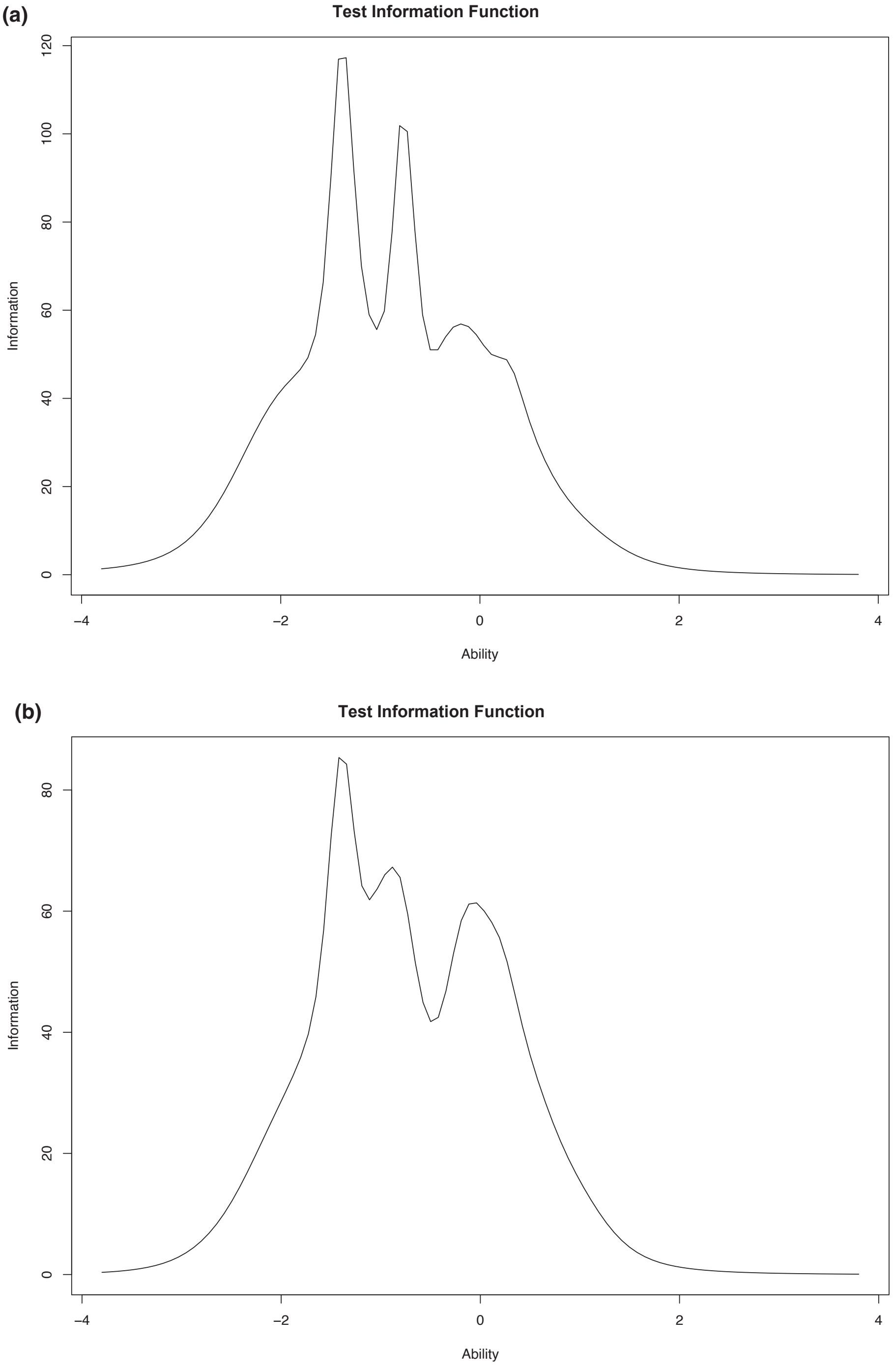

Fig. 2. The test information plot tells at which ability level most of the discriminative items were. In the initial test (Pane A), the highest points of the plot are centered around −1, which reflects that the initial test works best for discriminating between low-ability participants (i.e., L2 speakers). The curve after exclusion of the 17 easiest items is depicted in the Pane B. Note that the whole sample (L1 and L2 speakers) is depicted in this chart. The median vocabulary knowledge of the whole sample is centered around 0 in the *x*-axis.

$$\text{Score} = \text{Correct response to words} - (68/34)^{*}\ \text{Incorrect response to pseudowords}.$$

This means that if a participant responds to 50 out of 68 Finnish words correctly and 17 pseudowords erroneously his/her test score would be 50 − (68/34) * 17 = 50 − 34 = 16 points. This scoring procedure penalizes for guessing behavior, as: (1) a test taker who indicates most items to be words, even though several items seem relatively unfamiliar, will respond to a lot of pseudowords incorrectly leading to a very low score; (2) a test taker who indicates most items to be pseudowords, even though several items seem relatively familiar, will respond to a lot of words incorrectly which leads to a very low score as well. A participant who responds to each item at random – perhaps without much consideration – will respond to approximately half of the words and half of the pseudowords correctly and is expected to have a score around 0 (34–34 = 0). As it happens, test takers can even obtain a negative score if they get <50% correct on the words and/or pseudowords. Only someone who has all the words correct and does not respond to any of the pseudowords as a real word, gets the maximum score of 68.

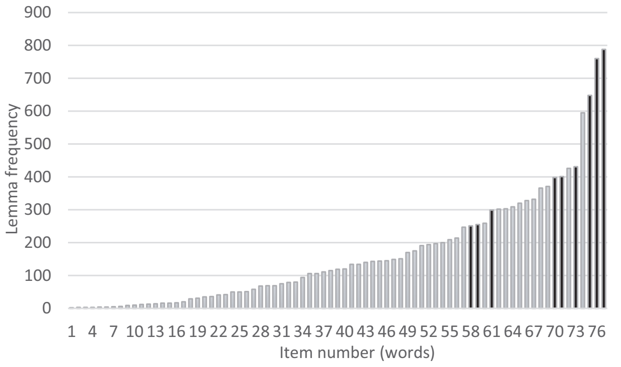

Fig. 3. An illustration of the distribution of the frequency of all words and the nine words that had lowest difficulty index in Item Response Theory (IRT) analysis (the excluded words are marked in black). This item distribution is based on those items that were included after CTT analysis.

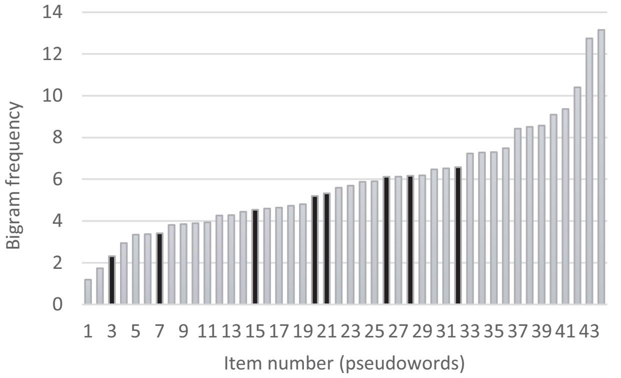

Fig. 4. An illustration of the distribution of the bigram values of all pseudowords and the eight pseudowords that had lowest difficulty index in Item Response Theory (IRT) analysis (the excluded pseudowords are marked in black). This item distribution is based on those items that were included after CTT analysis.

Table 3. *Lexical information of the final set of the items to be selected in Lexize*

	Words (n = 68)			Pseudowords (n = 34)		
	M	*SD*	Range	*M*	*SD*	Range
Length	5.76	1.10	4–9	6.12	1.28	4–9
Lemma frequency[a]	5.62	5.49	0.1–26.2	–	–	–
Bigram frequency[b]	5.94	2.52	1.5–12.7	6.16	2.73	1.2–13.2

[a]Scaled to 1 million.
[b]Scaled to 1,000.

RESULTS

Statistical analyses for the final item set

After the final set of items was established, we analyzed whether the Lexize score of the L1 group differed from that of the L2 group, and how our variables of interest correlated with Lexize scores within the two language groups. For the L1 group we assessed the correlation between Lexize score and age, school grade, and self-ratings. In the L2 group, we assessed the correlation between Lexize score and exposure (years lived in Finland) and self-ratings. We refrained from using regression analyses in this study, as explanatory data analyses are recommended over more complex predictive models when assessing the reliability of survey instruments (Shmueli, 2010). Spearman correlations were used for ranked data (i.e., school grades, self-ratings) and for continuous data that was skewed (age, exposure). The Wilcoxon rank-sum test for independent samples was used for comparing the L1 and L2 speakers, as their scores were not normally distributed. Similarly, the Kruskall–Wallis test was used to analyze whether there was a difference within the L1 group as a function of educational level (university level vs. secondary level vs. primary level). Post hoc tests for the non-parametric Kruskall-Wallis test were made with Dunn's multiple comparisons test (Siegel & Castellan, 1988).

Lexize scores for L1 and L2 speakers

The L1 group had a mean score of 60.8 (*SD* = 11.3; *Mdn* = 65). This was significantly higher than the mean score of the L2 speakers (*M* = 26.7; *SD* = 18.8; *Mdn* = 20); as indicated by the Wilcoxon rank-sum test for independent samples (*W* = 17,198, $p < 0.001$). The difference between L1 and L2 speakers is depicted in Fig. 5.

Correlations of Lexize score with Finnish school grade, self-ratings, age, and level of education in L1 speakers

A relatively high positive correlation was found between the Lexize score and school grade in Finnish language ($r_s = 0.57$, $p < 0.001$), reflecting that higher school grades are associated with better performance in Lexize. Note that the scale for Finnish school grades is 4–10. There was also a medium positive correlation for Lexize and self-ratings ($r_s = 0.41$, $p < 0.001$) showing that higher self-ratings are associated with higher Lexize scores. There was also a high correlation between Lexize scores and age ($r_s = 0.42$, $p < 0.001$), suggesting that increasing age is associated with improved performance. See Fig. 6 for illustration of these effects.

Moreover, performance in the Lexize test varied as a function of education level in the L1 group. The mean test scores were 65.7 (*SD* = 4.9), 61.2 (*SD* = 13.0), and 54.7 (*SD* = 11.0) for university level, secondary level, and primary level participants, respectively. Primary level refers to participants whose highest completed education level is comprehensive school (9 years of mandatory schooling in Finland), secondary level includes participants whose highest completed level is high school or vocational school, and university level includes participants with a bachelor's degree or higher. The Kruskall–Wallis test returned a significant main effect of educational level ($H = 42.36$, $df = 2$, $p < 0.001$). The Dunn's multiple comparison test showed that there was a difference in Lexize scores between primary and secondary level ($Z = -4.81$, $p < 0.001$), as well as between primary and university level ($Z = -6.31$, $p < 0.001$). The difference between secondary level and university level was close to significant ($Z = -1.88$, $p = 0.061$). See Fig. 7 for illustration of the effects and Table 4 for results of the Dunn's multiple comparisons test.

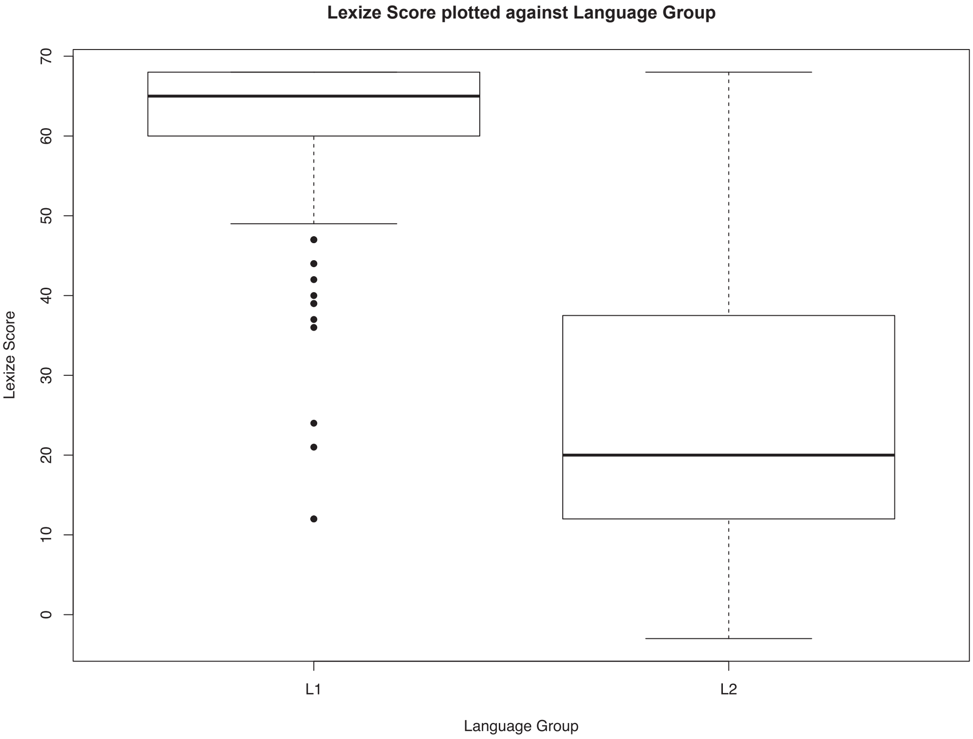

Fig. 5. Boxplots illustrating the differences in average Lexize vocabulary scores in the L1 and L2 group.

Correlations of the Lexize score with exposure and self-ratings in L2 speakers

In L2 speakers, the Lexize score correlated strongly with exposure ($r_s = 0.79$, $p < 0.001$) and self-ratings ($r_s = 0.70$, $p < 0.001$), reflecting that longer exposure to the Finnish language and higher self-ratings are associated with better performance in the Lexize test. See Fig. 8 for illustration of these effects.

DISCUSSION

In the current study, we developed and validated Lexize, an online vocabulary knowledge test in Finnish. The test was modeled after the English vocabulary test LexTALE (Lemhöfer & Broersma, 2012), which has been adapted for several other languages as well. Like the LexTALE tests, the Lexize test returned a clear difference between L1 and L2 speakers' vocabulary knowledge. This difference can be used as a proxy for difference in proficiency. Lexize was most accurate in estimating vocabulary knowledge among L2 speakers, as indicated by the test information curve showing that the test gives most information at the lower or middle level of the ability scale. Nevertheless, also among L1 speakers there was substantial variability, and this variability could be accounted for by certain L1 characteristics, as we will discuss below.

For L2 speakers, the Lexize scores correlated strongly with self-ratings, which is in line with earlier studies showing that vocabulary size is strongly associated with general communication ability (Laufer & Ravenhorst-Kalovski, 2010; Milton, 2010; Nation, 1993; Staehr, 2008). Vocabulary knowledge tends to show high correlations with both listening and reading comprehension measures (e.g., Zhang & Zhang, 2020) and also with writing and speaking (De Jong, Steinel, Florijn, Schoonen & Hulstijn, 2012; Miralpeix & Muñoz, 2018). Another important finding is that Lexize scores clearly correlate with exposure to Finnish, indicating that for the majority of L2 speaker's vocabulary size grows hand in hand with years spent in Finland. This in line with several earlier studies showing the impact of language exposure on vocabulary growth and general language proficiency development (Bowers & Vasilyeva, 2011; Grøver *et al.*, 2018; Herschensohn, 2009; Rydland *et al.*, 2014).

With regard to L1 speakers, the Lexize scores correlated moderately with Finnish school grades, indicating that better school grades in Finnish are associated with higher vocabulary knowledge. There was also a medium correlation between Lexize scores and self-ratings in Finnish, showing that participants with larger vocabulary size rate their native language skills somewhat higher than those with smaller vocabulary size. Additionally, there was a significant difference in Lexize scores between those who had completed primary education and those who had also completed secondary or university education. However, no significant difference was found between the latter two groups. It should be noted though that most of the L1 speakers that had completed their secondary education were enrolled at the university and will complete university education in due time. In sum, these results suggest that higher education level is related to higher test scores, but the test is not sensitive enough to pick up minimal differences at the higher end of education or then – after a certain point – lexical knowledge may even out. Finally, there was also a moderate correlation between age and Lexize score in the L1 group, suggesting that older age is associated with higher vocabulary level. However, the age range in the L1 group was not particularly large, so future studies will need to study this in more detail.

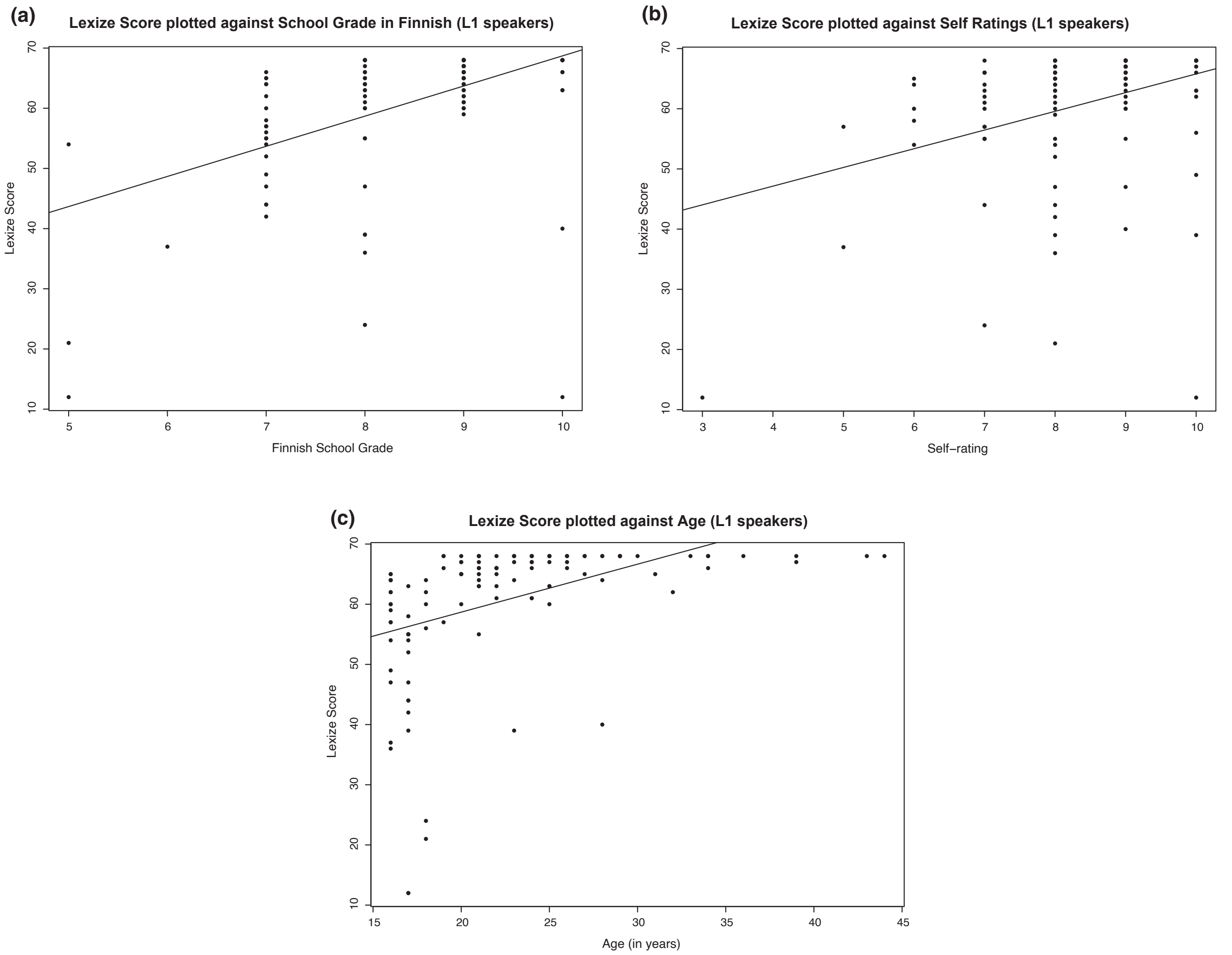

Fig. 6. Scatter plots and regression lines reflecting the positive correlations between Lexize scores and school grades for Finnish (A), self-ratings (B) and age (C) in L1 speakers.

In general, the results are in line with previous studies. That is, several L1 studies in other languages have reported similar correlations between test scores and school grade and/or self-ratings and/or age as well as a significant impact of education on vocabulary knowledge (Bowers & Vasilyeva, 2011; Brysbaert *et al*., 2016; Grøver *et al*., 2018; Herschensohn, 2009; Keuleers *et al*., 2015; Park *et al*., 2002; Rydland *et al*., 2014). The relationship between Lexize scores and the variables of interest is not as strong in the L1 group as in the L2 group, but the results show that Lexize is sensitive enough to reveal different levels of vocabulary knowledge among L1 speakers as well.

With respect to the test itself, the initial set of 132 items was reduced to 102 items through CTT and IRT analyses, mainly by excluding words from the two highest frequency bands with low to moderate discriminatory power. These analyses also revealed that the retained 102 items constitute an excellent set of items with high discriminatory power giving a good estimation of a person's vocabulary knowledge. As in previous studies, the difficulty of creating suitable pseudowords for L2 speakers also emerged in our study. This was confirmed by the high guessing rates for pseudowords in the IRT analysis. Brysbaert (2013) found that the lack of accents (e.g., *Bergere* vs. *Bergère*) and similarity in orthography (e.g., *oeiller* vs. *œiller*) blurred the boundary between words and pseudowords for L2 speakers of French and recommended restricting the use of items that differ from words in such a minimal way. Although there are no accents in Finnish, there are other idiosyncratic features in the Finnish phonological-orthographic system that may cause that pseudowords are formally close to real words. First, the phoneme-grapheme inventory in Finnish is relatively small, consisting of 13 consonants and 8 vowels. Especially the number of consonants is restricted, as the typical consonant inventory size in the world's languages is in the low twenties. This leads often to a dense phonological and orthographic neighborhood of Finnish words, and thus a substantial number of minimal pairs, especially when shorter words are concerned. In this case, if the pseudowords are created by changing only a few letters of the existing words, the pseudoword can still closely resemble several real Finnish words (e.g., *kukka*, *kakku*, *kokki*, *kukko*, *kokko* = real words; but *kokku* = a pseudoword). Lexical activation of the phonological neighbor may have caused the participants to be less certain about the correct answer (see e.g., Grainger, Muneaux, Farioli & Ziegler, 2005; Mathey, Robert & Zagar, 2004), and this uncertainty, in turn, would increase the guessing parameter index

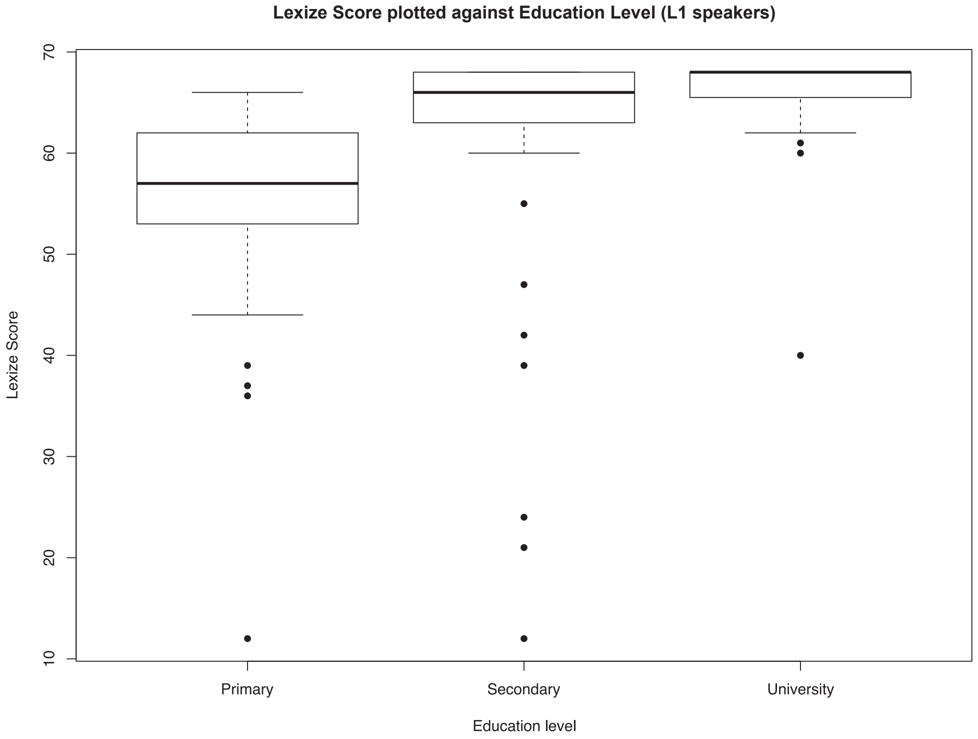

Fig. 7. Boxplots illustrating the differences in average Lexize vocabulary scores at different education levels in the L1 speakers.

Table 4. *Dunn's multiple comparisons test for education levels*

	Comparison	Z	P.unadj	P.adj
1	Primary–Secondary	−4.813024	<0.001	<0.001
2	Primary–University	−6.30697	<0.001	<0.001
3	Secondary–University	−1.876589	0.061	0.061

in the IRT analysis. However, as argued in Brysbaert (2013), it is important that the pseudowords and words do not differ too much from each other, as studies have shown that in that case it is possible to perform well in a word test without being proficient in the language involved (Grainger *et al.*, 2005; Keuleers & Brysbaert, 2010). Pseudowords that elicited a lot of mistakes for non-proficient but not for proficient L2 speakers of French were overregularizations, irregular words that were regularized (e.g., *metter* instead of *mettre* "to put"). Also in our vocabulary test, pseudoword items were created such that they differed from real words by one to three letters, while taking care that distinctions were not too minimal (pertaining to vowel or constituent length for instance). Most importantly in the current Lexize test, the final item set proved to be efficient in discriminating between the participants, so the possible resemblance between real words and pseudowords is not a critical problem in this study.

Limitations of the study

In the present study, Lexize was validated by comparing Lexize scores of L1 and L2 speakers and scores of L1 speakers of different educational levels, as well as by correlations between scores and self-ratings and years of exposure in L2 speakers, and Finnish language school grade in L1 speakers. We would like to note that especially self-ratings and school grades tap into a wider range of language skills than just vocabulary. The correlations of these variables with the Lexize scores thus imply that Lexize captures aspects of general language proficiency, that is, it underlines the notion that our vocabulary knowledge scores give an approximation of general language proficiency. However, for more definite conclusions as to whether Lexize scores reflect general language proficiency, we will need to validate the test against more extensive Finnish language tests like Kielo (Tani, 2008), which assesses several linguistic skills in different language modalities. We leave this to further studies.

Another issue that we like to take up is that Lexize, similarly to the different LexTALE tests, utilizes written word recognition and assesses vocabulary breadth rather than vocabulary depth. This format originates from Meara and Buxton's (1987) work which used a simple yes/no checklist for L2 speakers of English and showed that this type of test better predicts English examination results (Meara & Buxton, 1987) and is more accurate in a student placement test (Meara & Jones, 1988) than a vocabulary multiple-choice test. Ever since, several tests have consistently shown vocabulary breadth to be predictive of more general language skills, such as reading and listening comprehension (Zhang & Zhang, 2020). However, it should be pointed out that vocabulary depth is an important dimension of vocabulary knowledge as well. Several studies show that vocabulary depth can be disentangled from vocabulary breadth and that vocabulary depth independently explains variance in general language abilities like reading comprehension (e.g., Tran *et al.*, 2020) and writing production (Li

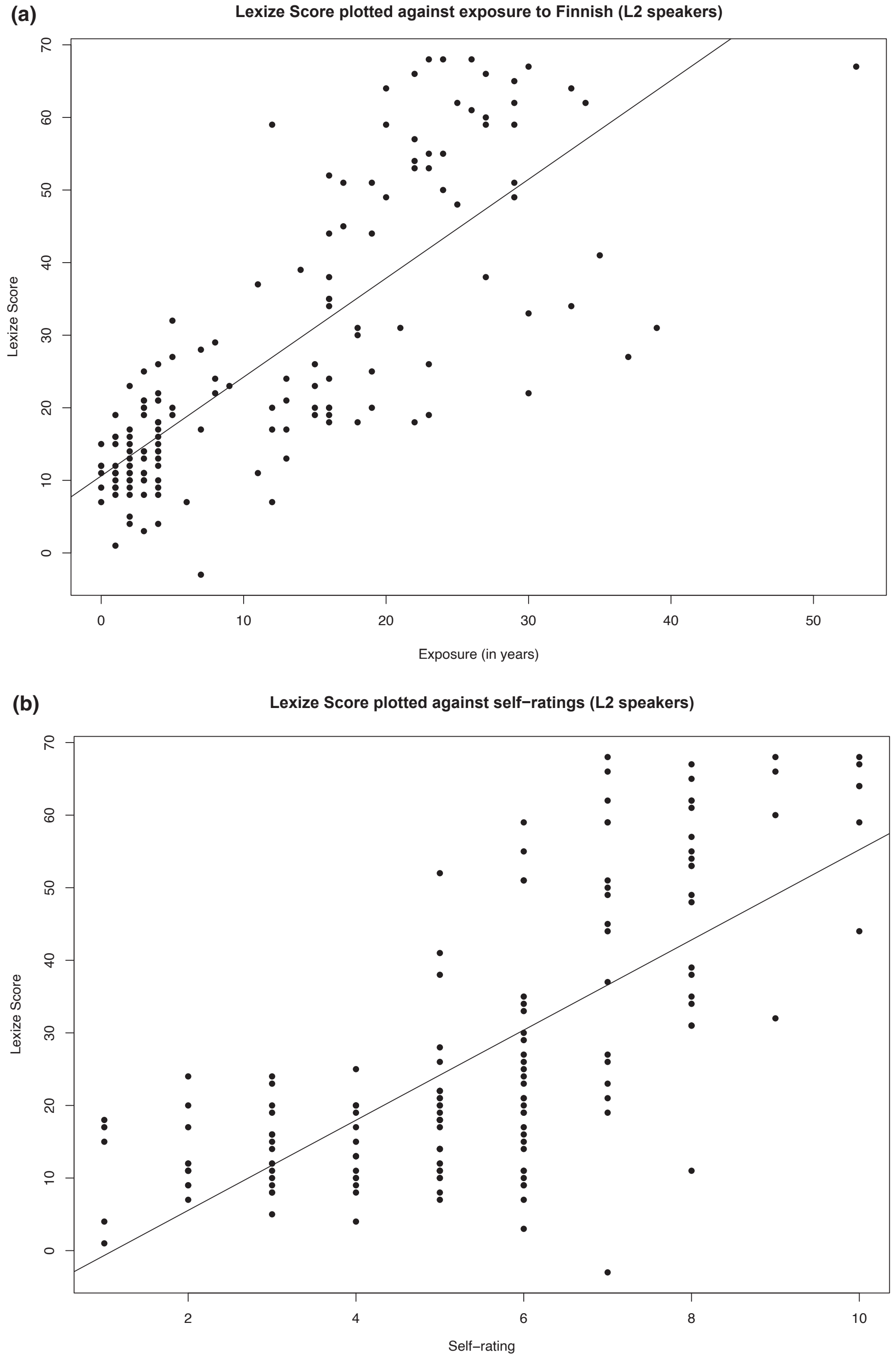

Fig. 8. Scatter plots and regression lines reflecting the positive correlations between Lexize scores and exposure to Finnish (Pane A) and self-ratings (Pane B) in L2 speakers.

& Kirby, 2014). Consequently, we assume that a vocabulary test like Lexize does not capture the full scope of vocabulary knowledge. We, however, leave it to future studies to investigate both dimensions of vocabulary knowledge and their relation to other linguistic abilities in more detail.

Conclusions and future directions

With the above-mentioned limitations kept in mind, we conclude that Lexize is a useful tool for research, as it provides a convenient and fast estimate of vocabulary knowledge and by that indicates general language proficiency in Finnish. In addition, as a similar test exists in English, French, Spanish, Chinese and Italian, it allows for cross-linguistic and bilingual research between Finnish and these languages using a relatively uniform measure of participants' vocabulary knowledge.

By virtue of the preceding questionnaire, Lexize can also provide interesting information about the factors that influence vocabulary knowledge and growth in learners of Finnish. In the current study, we have focused on exposure quantified by the number of years spent in Finland and age, but the questionnaire also includes questions for L2 speakers about L1 native language skills, percentage of daily use of Finnish, and motivation/importance to learn Finnish. The impact of these variables could be explored in future studies.

Apart from research, Lexize can be used to support L2 instruction. One possibility would be to use it for the estimation of baseline language proficiency of a participant before attending a language course or then to assess the development of the lexicon by comparing performance at different time points. This can be even done with L1 speakers at the earlier stages of education, as peak performance only seems to be reached during later adolescence. Finnish language users could also follow their vocabulary knowledge development themselves by means of Lexize. Moreover, Lexize could be used in clinical work, where measuring Finnish language proficiency is important when evaluating language impairments or following patient improvement in response to an intervention.

In short, Lexize provides a convenient and fast way to measure vocabulary knowledge in Finnish, which in turn can give an indication of a person's general proficiency level. It can be performed via our web-based test and can be completed within 5–10 min. Given that it is a reliable and free test, which is quick and simple to administer, we believe that Lexize will serve many researchers, educators and other people working in a domain where Finnish language skills are important to be assessed.

ACKNOWLEDGEMENTS

We are grateful to all participants who completed the Lexize test, as well as the teachers in various institutions in Finland, who kindly informed their students about the possibility to contribute to the study. We also would like to thank professor Mervi Eerola from the University of Turku department of statistics, for statistical consultation. This study is funded by the Minority Research Profile at Åbo Akademi University. The contribution of Raymond Bertram was within the project "Language and ethno-cultural variability of Southern Siberia in synchrony and diachrony: Language and culture interaction", RF government grant No. 14.Y26.31.0014.

The data that support the findings of this study are available from the corresponding author upon reasonable request.

NOTES

[1] Lexize can also be downloaded as a mobile application for iOS, Android, Microsoft operating systems via app stores, see https://psyk.abo.fi/Lexize/Lexize.html. However, in this study we provided only a link to the website, as we did not want to encourage repetitive use of the test in participants.

[2] Bigram frequency refers to the frequency with which adjacent pairs of letters (bigrams) occur in text, namely, letter strings having high bigram frequency have more common orthographical composition than those of low bigram frequency.

[3] According to Baker (2001), item discrimination is to be classified into the following categories: none 0; very low 0.01–0.34; low 0.35–0.64; moderate 0.65–1.34; high 1.35–1.69; very high > 1.70.

[4] According to Baker (2001), item difficulty is to be classified into the following categories: very easy < −2; easy −0.5,-2; medium −0.5,0.5; hard 0.5,2; very hard > 2.

REFERENCES

Amenta, S., Badan, L. & Brysbaert, M. (2020). LexITA: A quick and reliable assessment tool for Italian L2 receptive vocabulary size. *Applied Linguistics*, *42*, 292–314.

Baker, F. B. (2001). *The basics of item response theory* (2nd edn). College Park, MD: Eric Clearinghouse on Assessment and Evaluation.

Bates, E. & Goodman, J. (1999). On the emergence of grammar from the lexicon. In B. Macwhinney (Ed.), *The emergence of language* (pp. 29–79). Mahwah, NJ: Erlbaum.

Bowers, E. P. & Vasilyeva, M. (2011). The relation between teacher input and lexical growth of preschoolers. *Applied Psycholinguistics*, *32*, 221–241.

Brysbaert, M. (2013). Lextale_FR: A fast, free, and efficient test to measure language proficiency in French. *Psychologica Belgica*, *53*, 23–37.

Brysbaert, M., Stevens, M., Mandera, P. & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, *7*, 1116. https://doi.org/10.3389/fpsyg.2016.01116

Cameron, L. (2002). Measuring vocabulary size in English as an additional language. *Language Teaching Research*, *6*(2), 145–173.

Chan, L. & Chang, C. (2018). LEXTALE_CH: A quick, character-based proficiency test for Mandarin Chinese. *Proceedings of the Annual Boston University Conference on Language Development (BUCLD)*, *42*, 114–130.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory.* Austin, TX: Holt, Rinehart and Winston.

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R. & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*, 5–34.

Delgado, P., Guerrero, G., Goggin, J. P. & Ellis, B. B. (1999). Self-assessment of linguistic skills by bilingual Hispanics. *Hispanic Journal of Behavioral Sciences*, *21*, 31–46.

Gathercole, S. E. & Baddeley, A. D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language*, *28*, 200–213.

Grainger, J., Muneaux, M., Farioli, F. & Ziegler, J. (2005). Effects of phonological and orthographic neighbourhood density interact in visual word recognition. *The Quarterly Journal of Experimental Psychology*, *58*, 981–998.

Grøver, V., Lawrence, J. & Rydland, V. (2018). Bilingual preschool children's second-language vocabulary development: The role of first-language vocabulary skills and second-language talk input. *The International Journal of Bilingualism*, *22*, 234–250.

Gyllstad, H., Vilkaitė, L. & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL: Institut Voor Toegepaste Linguistik*, *166*, 278–306.

Hambleton, R. K. & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*, 38–47.

Herschensohn, J. (2009). Fundamental and gradient differences in language development. *Studies in Second Language Acquisition*, *31*, 259–289.

Hilton, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *The Language Learning Journal*, *36*, 153–166.

Honko, M. (2013). Alakouluikäisten leksikaalinen tieto ja taito. Toisen sukupolven suomi ja S1-verrokit [Lexical knowledge and skills in primary school children. Second generation L2 Finnish speakers and L1 peers]. Doctoral Thesis. Finland: University of Tampere, Tampere University Press. Retrieved July 28, 2021 from https://trepo.tuni.fi/handle/10024/94544

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

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

Keuleers, E. & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*, 627–633.

Keuleers, E., Stevens, M., Mandera, P. & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology*, *68*, 1665–1692.

Kimppa, L., Shtyrov, Y., Hut, S.C., Hedlund, L., Leminen, M. & Leminen, A. (2019). Acquisition of L2 morphology by adult language learners. *Cortex*, *116*, 74–90.

Laine, M. & Virtanen, P. (1999). *WordMill lexical search program*. Turku: Center for Cognitive Neuroscience, University of Turku.

Laufer, B. & Ravenhorst-Kalovski, G. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*, 15–30.

Lehtonen, M. & Laine, M. (2003). How word frequency affects morphological processing in monolinguals and bilinguals. *Bilingualism: Language and Cognition*, *6*, 213–225.

Lemhöfer, K. & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behaviour Research Methods*, *44*, 325–343.

Li, M. & Kirby, J. (2014). The effects of vocabulary breadth and depth on English reading. *Applied Linguistics*, *36*, https://doi.org/10.1093/applin/amu007.

Mair, P. (2018). *Modern psychometrics with R*. Cham: Springer International Publishing.

Mathey, S., Robert, C. & Zagar, D. (2004). Neighbourhood distribution interacts with orthographic priming in the lexical decision task. *Language and Cognitive Processes*, *19*, 533–560.

McLean, S., Kramer, B. & Stewart, J. (2015). An empirical examination of the effect of guessing on vocabulary size test scores. *Vocabulary Learning and Instruction*, *4*, 26–35.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjær & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 35–51). Cambridge: Cambridge University Press.

Meara, P. & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*, 142–154.

Meara, P. & Jones, G. (1987). Tests of vocabulary size as a foreign language. *Polyglot*, *8*, 1–40. Fiche 1.

Meara, P. & Jones, G. (1988). *Vocabulary size as a placement indicator (ED350829)*. ERIC. Retrieved July 28, 2021 from https://eric.ed.gov/?id=ED350829

Meara, P. & Miralpeix, I. (2016). *Tools for researching vocabulary*. Buffalo, NY: Multilingual Matters.

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. In I. Bartning, M. Martin & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, Second language acquisition and testing in Europe monograph series 1 (pp. 211–232). Rome: EUROSLA.

Miralpeix, I. & Muñoz, C. (2018). Receptive vocabulary size and its relationship to EFL language skills. *International Review of Applied Linguistics in Language Teaching*, *56*, 1–24.

Mochida, K. & Harrington, M. (2006). The yes/no test as a measure of receptive vocabulary knowledge. *Language Testing*, *23*, 73–98.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, *5*, 12–25.

Nation, I. S. P. & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*, 9–13.

Nation, P. (1993). Vocabulary size, growth, and use. In R. Schreuder & B. Weltens (Eds.), *The bilingual Lexicon* (pp. 115–134). Amsterdam: John Benjamins.

Niiranen, L. (2008). Effects of learning contexts on knowledge of verbs. Lexical and inflectional knowledge of verbs among pupils learning Finnish in Northern Norway. Unpublished doctoral dissertation. University of Tromsø. Retrieved July 28, 2021 from https://hdl.handle.net/10037/2109

Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N. S., Smith, A. D. & Smith, P. K. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology and Aging*, *17*, 299–320.

Perfetti, C. A. & Hart, L. (2001). The lexical basis of comprehension skill. In D. S. Gorfein (Ed.), *Decade of behavior. On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 67–86). Washington, DC: American Psychological Association.

Portin, M., Lehtonen, M. & Laine, M. (2007). Processing of inflected nouns in late bilinguals. *Applied Psycholinguistics*, *28*, 135–156.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved July 28, 2021 from https://www.R-project.org/

Raykov, T. & Marcoulides, G. (2010). *Introduction to psychometric theory*. New York: Routledge.

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal*, *19*, 12–25.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*, 1–25.

Rydland, V., Grøver, V. & Lawrence, J. (2014). The second-language vocabulary trajectories of Turkish immigrant children in Norway from ages five to ten: The role of preschool talk exposure, maternal education, and co-ethnic concentration in the neighborhood. *Journal of Child Language*, *41*, 352–381.

Saarela, L. (1997). Peruskoululaisten kirjoitelmien kehittyminen sanastotutkimuksen valossa [The developmental trajectory of elementary school children's essays in the light of vocabulary research]. Unpublished doctoral dissertation. Oulu: University of Oulu.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. New York: Palgrave Macmillan.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64*, 913–951.

Schmitt, N., Schmitt, D. & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing*, *18*, 55–88.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*, 289–310.

Siegel, S. & Castellan, N. J. Jr (1988). *Nonparametric statistics for the behavioral sciences* (2nd edn). New York: Mcgraw–Hill Book Company.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *The Language Learning Journal*, *36*, 139–152.

Tani, H. (2008). *Materiaalia aikuisten maahanmuuttajien suomen kielen taidon kartoitukseen ja kehityksen seurantaan*. Kielo, Finnish National Agency for Education. Retrieved July 28, 2021 from https://www.oph.fi/fi/tilastot-ja-julkaisut/julkaisut/kielo

Tran, A. H., Tremblay, K. A. & Binder, K. S. (2020). The factor structure of vocabulary: An investigation of breadth and depth of adults with low literacy skills. *Journal of Psycholinguistic Research*, *49*, 335–350.

Waller, M. I. (1989). Modeling guessing behavior: A comparison of two IRT models. *Applied Psychological Measurement*, *13*(3), 233–243.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*, 79–95.

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

Zimmerman, J., Broder, P. K., Shaughnessy, J. J. & Underwood, B. J. (1977). A recognition test of vocabulary using signal-detection measures, and some correlates of word and nonword recognition. *Intelligence*, *1*, 5–31.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Appendix S1. Lexical characteristics of the original stimulus set for real Finnish words. In bold the words that were excluded from the final version of Lexize.

Appendix S2. Background questionnaire of the Lexize test, available in English and Finnish.

Received 13 May 2020, accepted 29 June 2021