Behavior Research Methods (2022) 54:2625–2639
https://doi.org/10.3758/s13428-021-01731-1

LextPT: A reliable and efficient vocabulary size test for L2 Portuguese proficiency

Chao Zhou[1,2] · **Xinyi Li**[3]

Accepted: 18 October 2021 / Published online: 10 December 2021

Abstract

Vocabulary size has been repeatedly shown to be a good indicator of second language (L2) proficiency. Among the many existing vocabulary tests, the LexTALE test and its equivalents are growing in popularity since they provide a rapid (within 5 minutes) and objective way to assess the L2 proficiency of several languages (English, French, Spanish, Chinese, and Italian) in experimental research. In this study, expanding on the standard procedure of test construction in previous LexTALE tests, we develop a vocabulary size test for L2 Portuguese proficiency: LextPT. The selected lexical items fall in the same frequency interval in European and Brazilian Portuguese, so that LextPT accommodates both varieties. A large-scale validation study with 452 L2 learners of Portuguese shows that LextPT is not only a sound and effective instrument to measure L2 lexical knowledge and indicate the proficiency of both European and Brazilian Portuguese, but is also appropriate for learners with different L1 backgrounds (e.g. Chinese, Germanic, Romance, Slavic). The construction of LextPT, apart from joining the effort to provide a standardised assessment of L2 proficiency across languages, shows that the LexTALE tests can be extended to cover different varieties of a language, and that they are applicable to bilinguals with different linguistic experience.

Keywords L2 proficiency · Vocabulary size test · Bilingualism · Brazilian Portuguese · European Portuguese

Despite a considerable amount of research, the best way to assess second language (L2) proficiency accurately and reliably remains an ongoing question (Hulstijin, 2010; Leclercq & Edmonds, 2014). Therefore, different studies commonly resort to various methods to measure it (see Thomas, 1994 and Tremblay, 2011 for meta-analysis). The choice of assessment method is usually subject to the researchers' own understanding and to the feasibility (e.g. time limit) in the context of the given study.

In the experimental approach to L2 acquisition, wherein participants' L2 proficiency often needs to be assessed rapidly, many studies simply infer it on the basis of information collected through a background questionnaire. For instance, Tremblay (2011) analysed 91 L2 studies that had used a questionnaire-based assessment of L2 proficiency and reported that more than half (55) assigned participants to different proficiency levels conforming to years of L2 instruction (e.g. learners who had learned the target language for two years and six years, respectively) and institutional status (e.g. first- and second-year university students, respectively). This practice has, however, been criticised for being "hopelessly imprecise" (Hulstijin., 2010) because learners grouped in these criteria may differ dramatically in terms of cognitive ability, motivation, amplitude, and other factors that constrain L2 acquisition. These differences might lead to highly dissimilar paces of L2 development. In other studies adopting the questionnaire approach, participants were asked to rate their own L2 ability on a scale (e.g. a ten-point Likert scale, in which 1 represents the lowest and 10 the highest) or according to some predetermined categories (e.g. "beginning", "intermediate", and "advanced"). Although this kind of self-assessment is quick and can provide some insight into learners' L2 proficiency (Oscarson, 1989; LeBlanc & Painchaud, 1985), its validity has been shown to vary across studies (see Marian et al. 2007 for a review). The inconsistency of self-assessment may reflect the fact that it is subject to many factors, including the wording of the questions, the

✉ Xinyi Li
xinyi.li@campus.fcsh.unl.pt

[1] University of Lisbon, Lisbon, Portugal

[2] University of Minho, Braga, Portugal

[3] NOVA University of Lisbon, Lisbon, Portugal

language skills being assessed (reading, writing, speaking, and listening), the proficiency level of the students, and even the cultural background of the participants (Strong-Klause, 2000).

The lack of reliability in the questionnaire-based approach has led to many attempts to create a more objective assessment of L2 proficiency. For instance, some studies asked participants to report the scores that they had obtained in a standardised proficiency test (Tremblay, 2011), such as the International English Language Testing System (IELTS) or Test of English as a Foreign Language (TOEFL) for L2 English. These standardised tests have generally been validated in various ways over the years and thus provide more precise and reliable information on participants' L2 proficiency. Nevertheless, this method might hinder subject recruitment in experimental research, since it is necessary to ensure that all participants have recently taken a certain standardised test (otherwise, learners' actual proficiency might not be reflected). Naturally, a better alternative would be to administer an existing standardised test as part of the experimental study. Yet, while this is a more reliable method in comparison with the aforementioned methods, the feasibility is quite restricted. The fact that standardised tests are usually quite costly and time-consuming complicates their integration into experimental research, where participants are very often asked to perform several tasks. Therefore, an efficient, reliable, and open-access L2 proficiency assessment tool is desirable.

Assessment of proficiency levels and lexical knowledge

Studies aiming to construct an efficient L2 proficiency test have focused on several features that may serve as reliable indicators of a learner's L2 ability, among which vocabulary size has attracted considerable attention (Milton, 2013; Nation, 2013). Adequate lexical knowledge is viewed as the prerequisite of effective language use. It has been evidenced that the breadth of L2 lexical competence grows as a result of an increase in language proficiency (Meara, 1996; Bonk, 2000; Zavera et al., 2005). Moreover, a high correlation between proficiency testing and vocabulary testing has been reported in many studies (e.g. Qian, 1999; Beglar & Hunt, 1999; Nizonkiza, 2011). These results together suggest that lexical competence is a reliable predictor of L2 proficiency.

A solid indicator of lexical competence is the receptive vocabulary size since it provides useful information on how vocabularies develop (Eyckmans, 2004). The receptive use of vocabulary, according to Nation (2001), essentially involves the ability to perceive the form of a word and retrieve its meaning while listening or reading. A quite extensive body of research evidence indeed supports the idea that receptive vocabulary size is a reliable indicator of overall L2 proficiency. In particular, these studies show that vocabulary size correlates with all four main elements normally assessed in a standardised language test, i.e. reading comprehension (Beglar & Hunt, 1999; Laufer, 1992; Qian, 1999; Stæhr, 2008), writing ability (Astika, 1993; Laufer, 1998; Stæhr, 2008), listening comprehension (Milton et al., 2010; Stæhr, 2008; Zimmerman, 2004), and oral fluency (Milton et al., 2010; Zimmerman, 2004). To sum up, converging evidence in the literature suggests that a vocabulary size test can be regarded as a sound instrument to assess overall L2 proficiency.

LexTALE and its equivalents in different languages

Faced with the need for a quick, valid, and accessible tool to assess L2 English proficiency, Lemhöfer and Broersma (Lemhöfer & Broersma, 2012) developed the LexTALE (Lexical Test for Advanced Learners of English; available at http://www.lextale.com) vocabulary test.

The LexTALE test takes approximately 3.5 minutes to complete. It consists of 60 items in total, of which 40 are words and 20 are nonwords. The test items were selected from an unpublished vocabulary size test introduced by Meara (1996) in such a way that real words span various frequency tiers. It is expected that low-frequency words should only be known to L1 and highly advanced L2 speakers, whereas high-frequency words should be recognised by learners of all proficiency levels. Nonwords were included in order to militate against response bias, i.e. identifying unknown words as real words, and the 2:1 real-to-nonce ratio is in accordance with classical yes–no vocabulary tests (e.g. Meara & Buxton, 1987; Meara, 1992).

For validation, the assembled LexTALE test, a translation task, and a commercial standard test (Quick Placement Test 2001, hereinafter: QPT) were administrated to two groups of L2 English learners, 72 L1 Dutch speakers and 87 L1 Korean speakers (the Korean participants were also asked to report their TOEIC[1] scores), who also completed a self-ratings and language background questionnaire. Results showed that the LexTALE scores were more closely correlated with participants' performance on the translation task and with their TOEIC scores compared to the self-ratings. In the case of the performance on the QPT, the LexTALE scores corresponded better with the QPT scores and manifested a much lower

[1] The Test of English for International Communication (TOEIC) is a standard English proficiency test for non-native speakers.

false alarm rate[2] than the self-ratings did. These results suggest that, apart from its quick implementation, LexTALE is more accurate than self-assessment in terms of reflecting actual L2 proficiency, screening participants, and selecting who truly reaches a certain proficiency prerequisite for the forthcoming experiment.

The validity of LexTALE has been further evidenced by psycholinguistic research. For instance, in a word recognition study by Diependaele et al. (2013), LexTALE scores successfully accounted for the difference in the size of the word frequency effect within and between L1 and L2 groups. Moreover, Khare et al. (2013) reported that the magnitude of attentional blink is significantly and strongly correlated with the proficiency levels assessed by LexTALE, i.e. there is a larger attentional blink effect for more highly proficient bilinguals.

The efficiency and reliability of the LexTALE test have led to its extension to other languages in the last few years. Building on the original LexTALE, the subsequent tests have introduced some modifications and innovations for cross-linguistic implementation. The French equivalent, LexTALE_FR (Brysbaert, 2013), started off with lexical databases collected from written sources and film subtitles and from lexical decision tasks, in such a way that selected test items are more representative of participants' linguistic exposure in real life (Brysbaert & New, 2009). In an attempt to extend the test to L1 speakers, the author increased the number of test items to 84 (56 words, 28 nonwords) from different frequency levels allowing a better coverage of the whole proficiency range. In terms of the test format, the French test replaced the original yes–no template with a checklist (a go/no-go task) due to concerns that the first method could be demotivating. The checklist format was also adopted by the following versions of LexTALE. The Spanish version Lextale_Esp (Izura et al., 2014) further improved the quality of the test by starting with a large scale of 180 items (90 words and 90 nonwords) and selected the most suitable 90 items (60 words and 30 nonwords) on the basis of the pilot results. The same scale of 90 items was later adopted in the Chinese (LEXTALE_CH; Chan & Chang, 2018) and Italian (LexITA; Amenta et al., 2020) extensions. In the case of LEXTALE_CH, the test was extended to a language with a logographic writing system. LexITA further assessed the test validity by comparing the test scores with participants' CEFR[3] proficiency levels. Despite all the differences, all extensions have demonstrated robust validity and consistency, indicating the potential of the cross-linguistic extension of LexTALE.

However, it is worth noting, as pointed out in Izura et al. (2014) and Amental et al. (2020), that validation studies of previous extensions of LexTALE were conducted with a rather homogeneous group of L2 learners, i.e., a good proportion of them spoke the same L1 and were students of the same or similar institutions. The remaining question is whether LexTALE tests are appropriate for learners with different L1 backgrounds, especially those with L1s typologically approximate to the L2. Ferré and Brysbaert (2017) tackled the issue of proximity by testing Spanish-Catalan bilinguals and showed that the Spanish-dominant group outperformed the Catalan-dominant one on Lextale_Esp. This provides evidence that the sensitivity of the test is not mitigated by the typological similarity.

Method

Developing a Portuguese extension of LexTALE

Expanding on the standard procedure of test construction adopted in previous LexTALE tests, in the current study, we develop an objective and easy-to-use vocabulary size test for L2 Portuguese proficiency, named LextPT. As the LexTALE test and its extensions, LextPT allows quick and easy administration and integration into experimental research. As a standardised test, it makes it easier to compare results obtained in different studies. More importantly, the estimate of L2 Portuguese proficiency is subsequently represented on a continuous scale of test scores, allowing researchers to gain insight into individual differences in language processing and acquisition (see Diependaele et al., 2013).

As it is intended to cover the representative population learning Portuguese as an L2, the construction and validation of LextPT took into account both the European and Brazilian varieties[4]. Although being two variants of the same language, European and Brazilian Portuguese differ both in terms of grammar (e.g. phonology, morphology, syntax and semantics) and lexicon (see Wetzels et al. 2016 for an overview). Pertaining to the lexical differences, previous studies suggest that about 11% of the general lexical items (not specific to small communities or to technical subjects) have contrastive use between the European and Brazilian varieties

[2] The percentage of subjects selected for participation under the prediction, but who did not actually obtain the minimum QPT score required.

[3] The Common European Framework of Reference for Languages (CEFR), led by the European Council and launched in 2001, is a worldwide standard for organiszing foreign language proficiency in six levels (A1, A2, BA, B2, C1, C2), currently available in 40 European and non-European languages.

[4] Portuguese is the language of over 230 million people, about 15 million of whom are speakers of European Portuguese (Segura, 2013), and more than 170 million of whom are speakers of Brazilian Portuguese (Mattos e & Silva, 2013). However, it is difficult to accurately count the number of speakers of the Angolan, Mozambican, Cape Verdean, Guinean, São Tomean, Timorese, and Galician varieties.

(Wittmann et al., 1995; Barreiro et al. 1996). For instance, some words are used in both variants but with semantic differences (European: *banheiro*, "lifeguard"; Brazilian: *banheiro* "washroom") and some words only differ in spelling (European: *linguista*; Brazilian: *lingüista*, "linguist"). Therefore, the considerable overlap of general lexical items leads us to believe that it is feasible to construct a vocabulary size test that can be administrated equally well to learners of both Portuguese varieties.

In the rest of this section, we describe the development of LextPT. The procedure of item selection closely observed the criteria adopted in LexTALE and its extensions, especially the Spanish (Izura et al., 2014) and Italian (Amenta et al., 2020) versions. The construction was composed of two studies, a pilot study to carry out the item selection, and a validation study to test the validity of the selected items.

Material

Following previous extensions of LexTALE, the development of LextPT started off with a total of 180 items, comprising 90 real words and 90 nonwords. We consider that such a number of items suitable for being integrated into experimental research while covering a broader range of proficiency.

Real word items were extracted from two subtitle-based lexical databases, using word frequency as the selection criterion. As for language register, in comparison with corpora drawn from written sources, e.g. books, magazines and newspapers, usually edited or polished, word frequency measured on the basis of television and film subtitles has been shown to be more representative of the spontaneous language use and daily linguistic exposure of the population frequently recruited in psycholinguistic experiments (Brysbaert & New, 2009). The item selection began with the SUBTLEX_PT[5] database (Soares et al., 2015). All items in SUBTLEX_PT were divided into six frequency tiers (per million words) in accordance with prior research (Brysbaert, 2013; Izura et al., 2014; Chan & Chang, 2018; Amenta et al., 2020). Only nouns and adjectives were considered, while compounds and derived words were left out of the selection. After selecting the first version of 90 real word items, we turned to a homologous lexical database of Brazilian Portuguese, SUBTLEX_PT_BR[6] (Tang, 2012) to check whether the selected items belonged to the same frequency interval (items in SUBTLEX_PT_BR were likewise subdivided into six intervals). Items that did not fit into the same frequency level in both corpora were excluded and we continued to evaluate novel candidates. This was repeated until all 90 items were roughly equivalent in terms of frequency in both corpora. These 90 selected items varied from highly frequent words, probably recognizable for L2 beginners, such as *música* "music", *razão* "reason", *máquina* "machine", to very low-frequency words that should be known to only proficient native speakers or highly advanced learners, such as *fatídica* "ominous", *espólio* "spoils", and *jusante* "downstream". The majority of the 90 word items are nouns ($n = 57$), followed by adjectives ($n = 19$) and items that can belong to both classes ($n = 14$). The distribution of selected items in terms of frequency (occurrences per million words; pm) is shown in Table 1. The selected items were skewed towards low-frequency tiers, with the purpose of having items with different difficulty levels, simultaneously increasing the overall difficulty of the test. Consequently, LextPT can cover a wide range of proficiency and effectively discriminate among advanced L2 learners. The spelling of all selected items follows the Portuguese Language Orthographic Agreement of 1990, a unified orthography signed or later adhered to by all the countries that have Portuguese as their official language.

Table 1 Distribution of real word items across frequency tiers

Frequency tier (pm)	Number of items
<1	26
1–5	23
6–10	14
11–20	17
21–100	8
> 100	2

The 90 nonwords were adopted from two existing stimuli lists (Justi et al., 2014; Venâncio, 2018), whereby all nonword items resembled the Portuguese ortho-phonotactic structure[7]. In this way, participants were expected to rely only on their lexical knowledge for judgment, instead of the structural well-formedness of Portuguese. The average OCD20 (Levenshtein distance 20) value of the selected nonwords is around 2. In other words, they do not have an extensive number of orthographic neighbours (thus no great proximity to the lexicon). This is important because although LextPT is not a timed test and the participants were informed of this in the instructions, it is worth preventing participants from regarding a

[5] SUBTLEX_PT is a lexical database containing 132,.710 Portuguese words, obtained from a 78-million-word corpus based on subtitles of European Portuguese film and television series screened between 1990 and 2011.

[6] SUBLEXT_PT_BR comprises 136,.147 word types obtained from 61 million words of conversational Brazilian Portuguese.

[7] The nonword items are in fact all pseudo-words. But in line with previous LexTALE tests, we refer to them here as nonwords.

nonword item as "real" out of negligence stemming from a mere glance at the item. According to Brysbaert (2013), some of the nonwords he selected proved to be inadequate for the test as they elicited more errors in the responses by native French speakers than in those by learners because they were pseudo-homophones of very low-frequency words, part of fixed expressions, or derived from real words in the absence of a proper graphic accent. This possibility was taken into account, and nonword items with the above characteristics were excluded. Moreover, we also excluded items without a very limited OLD20 value yet orthographically similar to high-frequency words, such as *chiança* (*criança,* "child"), *bolanço* (*balanço,* "balance"), in order to avoid errors caused by negligence. Finally, we conducted a search of the selected nonwords in an *online* dictionary, *o Dicionário Priberam da Língua Portuguesa* (https://dicionario.priberam.org/), and in the search engine Google to ascertain that these were not existing words in Portuguese with a low frequency or existing results of neological creation.

Procedure

The pilot study was set up using Google Forms. The link to the questionnaire was shared through social media and distributed to a mailing list of teachers of Portuguese as an L2.

At the beginning of the questionnaire, participants were given options to read the upcoming questions and instructions in English or Portuguese. The questionnaire consisted of two parts. The first part included a consent form for participation and questions regarding the participants' sociolinguistic background, i.e. native language, age, gender, other languages they spoke; a self-assessment of their Portuguese overall proficiency on a ten-point Likert scale ranging from 1 (lowest) to 10 (highest); and how many years they had been learning Portuguese. Upon providing this information, the participants were given detailed instructions concerning a pilot lexical task. Specifically, they were asked to indicate which Portuguese words they knew or believed to be real Portuguese words, even if they were not sure of their exact meaning. The second part of the questionnaire was the lexical task comprising the 180 items, which were arranged into a semi-randomised presentation to ensure that no more than five real words or nonwords appeared in succession (Lemhöfer & Broersma, 2012). The same item presentation order was shown to all participants. It was made clear that the test was anonymous, and that it was to be completed individually and without consulting other people or dictionaries.

Participants

L1 and L2 speakers of EP and BP were recruited in the pilot study in order to select the most suitable items for assessing the vocabulary size of both European and Brazilian varieties. 19 speakers from the L1 group (10 EP and 9 BP) were excluded in the following analysis because they claimed to have not grown up monolingually. Thus, the L1 group consisted of 130 participants in total: 69 native speakers of EP (hereafter, "L1 EP"; 55 women, 14 men, mean age = 32.63 years, SD = 12.6) and 61 speakers of BP (hereafter, "L1 BP"; 30 women, 29 men, 2 others, mean age = 34.85 years, SD = 12.9). Of these 130 L1 participants, 118 also spoke one or more languages apart from Portuguese, while 12 gave no answer to this question.

The L2 group was composed of 120 participants, 71 learners of EP (hereafter, "L2 EP"; 54 women, 15 men, 2 NA, mean age = 28.84 years, SD = 9.56) and 49 learners of BP (hereafter, "L2 BP"; 33 women, 14 men, 2 others, mean age = 30.11 years, SD = 13.97). All these participants whose responses were included in data analysis are late L2 learners of Portuguese. The L1s of these participants are listed in Table 2.

Table 2 Pilot study: L2 group participants' native languages

L1	Number of participants	
	L2 EP	L2 BP
Chinese	39	37
Italian	19	2
Spanish	-	6
English	3	2
Bilingual English/Spanish	2	-
Armenian	2	-
French	1	1
Korean	-	1
Romanian	1	-
Dutch	1	
Ukrainian	1	
Russian	1	
Slovakian	1	

Results

The first part of this section outlines the item selection procedure to be integrated into the final version of LextPT, which consists of only 60 word items and 30 nonword items.

Selecting items for LextPT

Two items were excluded initially: a word item *oxigénio*, because EP and BP differ with respect to the use of accent

mark in this word (PT: *oxigénio* and BP: *oxigênio*)[8], and a nonword item *elvidi* because all native and non-native pilot participants rejected it as a real Portuguese word.

The quality of the remaining 178 items was first examined using point-biserial correlation. The point-biserial correlation analysis, which evaluates the relationship between the response to each item and the participants' total accuracy, sheds light on the usefulness of an item. This correlation ranges between −1 and +1: a positive correlation indicates that participants who have high overall test scores tend to perform better on the given item than those who have relatively low scores, while a negative correlation suggests an anomalous situation in which participants with high overall scores perform less well on this item than those with low scores. In other words, the most important information provided by this analysis is that good items should not give rise to a negative value.

The point-biserial correlation was performed on the word and nonword items separately using the *ltm* package (Rizopoulos, 2006) in R (R Development Core Team, 2020). The responses of both the L1 and L2 speakers were included in the analysis. A positive correlation was found for all word items (from 0.02 to 0.81) as well as for all nonword items (from 0.23 to 0.66). For the final version of LextPT, we intended to include those items that equally span a wide range of difficulty levels and, at the same time, maintain good discrimination power. Hence, we further examined these 178 items in an item response theory (IRT) analysis, which takes into consideration both items' difficulty levels and discrimination power. Discrimination power refers to how well an item can distinguish a more proficient participant from a less proficient one. The IRT analysis was performed on word items and nonword items separately, also using the *ltm* package. Based on the results of the IRT analysis, we ordered the items conforming to difficulty levels (Izura et al., 2014), divided them into 30 approximately equal intervals, and then extracted from each interval those items with the best discrimination power. An illustration of the IRT analysis can be found in Fig. 1, where the x-axis represents the difficulty level (highest difficult level: 4), and the steepness of the response curve in its middle section reflects an item's discrimination power (i.e. a steeper curve signals a stronger discrimination power). In particular, as shown in Fig. 1, the word *jusante* "downstream" is more difficult than the words *tenro* "tender" and *nupcial* "nuptial"; we can also see that *nupcial* holds more discrimination power than *tenro*, despite their similar difficulty levels.

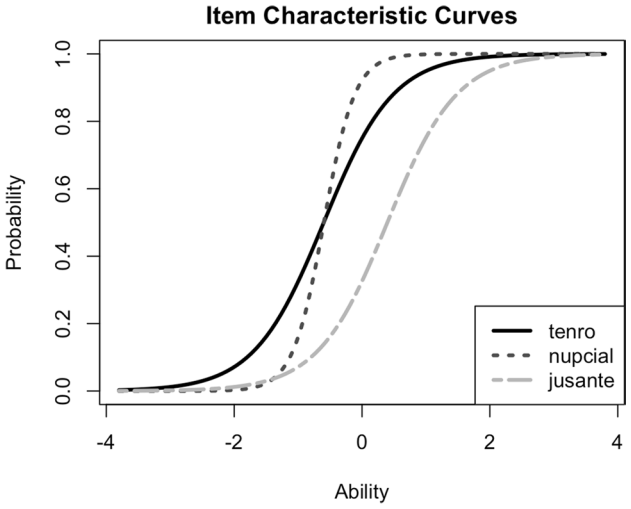

Fig. 1 Item response curves for three word items illustrating the item selection procedure for LextPT

A total of 90 items (60 real words and 30 nonwords) were selected for the LextPT test. Their characteristics are reported in Table 3.

Scoring of LextPT and reliability analysis

In line with prior research (Lemhöfer & Broersma, 2012; Brysbaert, 2013; Izura et al., 2014; Chan & Chang, 2018; Amenta et al., 2020), the LextPT test score was computed according to the following equation, which penalises guessing behaviour (e.g. randomly selecting words).

$$\text{LextPT Score} = N_{\text{yes to words}} - 2 \times N_{\text{yes to nonwords}}$$

The maximum score of 60 can only be achieved if someone identifies all real words and does not select any nonword. The LextPT scores were calculated for all 250 pilot participants, and their results are summarised in Table 4.

The LextPT scores indicate that the L2 group performed substantially less well than the L1 group [Welch-corrected two-sample $t(137.8) = 19.265$, $p < .0001$], with a large effect size [Cohen's $d = 2.52$]. This difference is in line with that observed by Brysbaert (2013) in the French test, Izura et al. (2014) in the Spanish test, and Chan and Chang (2018) in the Chinese test.

The LextPT scores of the 120 Portuguese learners were first compared with their self-assessment proficiency scores. In view of prior research, we expected to observe a moderate correlation because self-assessment has been shown to reflect actual L2 proficiency to some extent, although it is not perfect (Marian et al., 2007; Brysbaert, 2013; Izura et al., 2014; Chan & Chang, 2018; Amenta et al., 2020). After correlating L2 participants' LextPT scores with their

[8] Although both orthographic forms, oxigénio and oxigênio, can be found in SUBTLEX-PT-BR (Tang, 2012), 14 out of 61 L1-BP speakers rejected oxigénio as a real Portuguese word.

Table 3 Characteristics of word and nonword items included in the final version of LextPT

Distribution	Words				Nonwords	
	No. letters	No. syllables	Frequency (per million)		No. letters	No. syllables
			PT	BR		
Min	5	2	0.0385	0.0667	5	3
Max	9	4	167.740	241	8	3
Mean	7.217	3.217	13.181	14.904	6.5	3
SD	1.121	0.555	27.185	34.782	0.682	0

Table 4 Summary of the LextPT results from the pilot study

	L1 EP	L1 BP	Overall L1	L2 EP	L2 BP	Overall L2
Mean	54.62	55.26	54.92	30.62	27.80	29.46
SD	4.79	3.10	4.08	15.62	10.98	13.93
Range	30–60	46–60	30–60	0–58	−1 to 53	−1 to 58

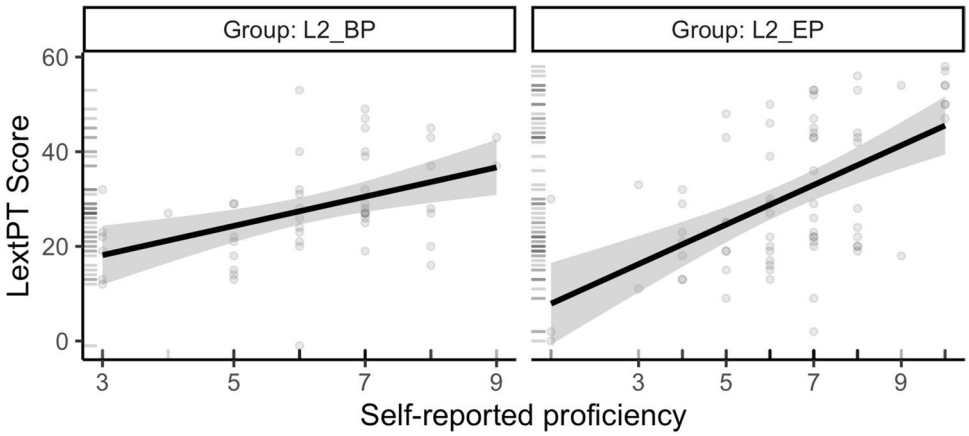

Fig. 2 LextPT scores (corrected accuracy; maximum 60) by self-rated Portuguese proficiency (1–10 scale; maximum 10) in the pilot study

self-assessment ratings (range = 1 to 10), a moderate, statistically significant correlation was indeed found for both the L2 EP group [Pearson's $r(69) = .56$, $p < .0001$] and the L2 BP group [Pearson's $r(47) = .45$, $p < .01$], see Fig. 2. As pointed out by Brysbaert (2013), the lack of reliability of self-assessment can be attributed to the fact that many participants' reference on language proficiency hinges on a rather narrow group. For example, L2 beginners tend to compare themselves with their novice peers, and consequently, as long as an L2 beginner thinks that she outperforms at least half of her fellows, she may rate herself as 6–8, regardless of her still limited vocabulary size; on the other hand, an advanced learner may have the tendency to evaluate her L2 proficiency in comparison with native speakers and, accordingly, also rate herself as 6–8.

Another correlation analysis was performed between the LextPT results and years that the learners had spent learning Portuguese. Although a moderate and significant correlation is present for both L2 EP group [Pearson's $r(66) = .52$, $p < .0001$] and L2 BP group [Pearson's $r(45) = .42$, $p < .01$], there is a considerable degree of dispersion in the data, as illustrated in Fig. 3. For both L2 groups, some learners who reported having learned Portuguese for more than 10 years were outperformed by those who had learned it for 4–6 years. These results corroborate with what was found for French learners in Brysbaert (2013), suggesting that, as an L2 proficiency assessment method, learning length cannot be sensitive to individual variation within a group of participants who have spent a similar amount of time acquiring the language.

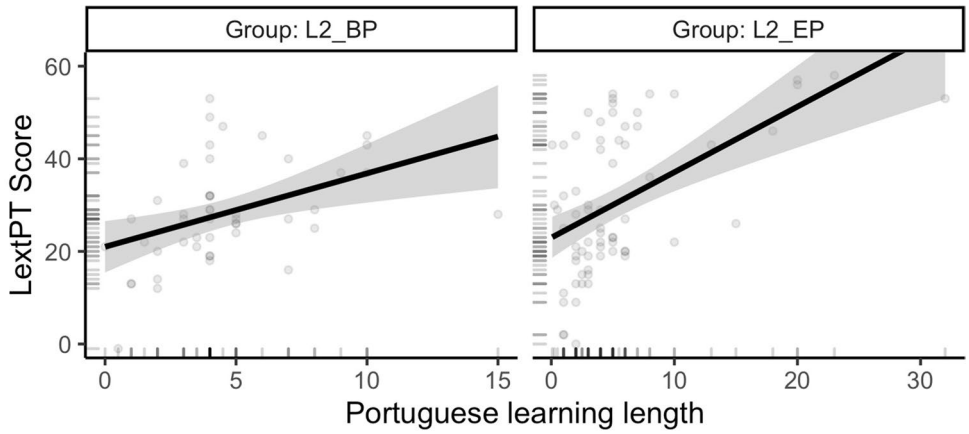

Fig. 3 LextPT scores (corrected accuracy; maximum 60) by Portuguese learning length (years) in the pilot study

The last analysis performed on the pilot results was a reliability test. Following Amenta et al. (2020), the internal consistency was measured with Cronbach's alpha and the ICC coefficient using the *psych* package (Revelle, 2021). The reliability of LextPT turned out to be excellent, considering both L1 and L2 participants [$\alpha = .97$; ICC3k = .96, $p < 0.001$]. This high reliability remains true for both the L2 EP group [$\alpha = .93$; ICC3k = .93, $p < 0.001$] and the L2 BP group [$\alpha = .88$; ICC3k = .88, $p < 0.001$], suggesting a high internal consistency in the performance of learners of both European and Brazilian varieties when measured with LextPT. In order to make sure that LextPT is suitable for learners from different L1 backgrounds, we conducted another reliability analysis on two L2 subgroups of learners, regardless of the variety that they were learning (or: whether they were learning BP or EP), whose L1s are respectively Chinese and Romance languages (Italian, Spanish, French, Romanian). Chinese and Romance languages are typologically distinct from each other on several parameters in the sense of Comrie (1989), e.g. word order (of the relative clause and the head noun), case systems, morphological typology, tone. The results confirmed that LextPT is very reliable for both Chinese speakers [$\alpha = .88$; ICC3k = .88, $p < 0.001$] and Romance language speakers [$\alpha = .93$; ICC3k = .93, $p < 0.001$].

Testing the final version of the LextPT

In the pilot study, the responses of 250 participants were evaluated using point-biserial correlation and IRT analysis, according to which 60 word items and 30 nonword items were selected. These 90 items, which span various difficulty levels and have the best discrimination power, were included in the final version of LextPT. The reliability analysis indicated that LextPT is not only satisfactory for learners of both European and Brazilian varieties, but also appropriate for participants from different L1 backgrounds.

However, recall that the responses to the 90 items included in LextPT were elicited together with other items that were later excluded; this may have had an impact on the responses. Moreover, even though the presentation of the initial 180 items was semi-randomised, the same order of presentation was applied to all participants (as in Lemhöfer & Broersma, 2012; Brysbaert, 2013; Izura et al., 2014; Chan & Chang, 2018), which might have given rise to effects of list composition (Amenta et al., 2020). We carried out a validation study to check the quality of the final items in the absence of the excluded items. The format and administration of the validation test were identical to those in the pilot study, except that only the selected 90 items were used and the item presentation was randomised for each participant. In the validation phase, L2 participants were further asked to provide information on the levels of their attained CAPLE[9] or CELPE-Bras[10] certificates. If a learner had never obtained a certificate on Portuguese proficiency but was taking a Portuguese language course at the moment of participation, she was instructed to indicate the level of that course.

[9] CAPLE (Centro de Avaliação de Português Língua Estrangeira or Centre for Evaluation of Portuguese as a Foreign Language) exams, developed by the University of Lisbon, aim at certifying the proficiency of European Portuguese as a foreign language, offered at six reference levels, from A1 (beginner) to C2 (near-native), conforming to the Common European Framework of Reference of Languages (CEFR). For detailed information, please consult: https://caple.letras.ulisboa.pt/pagina/1/caple

[10] CELPE-Bras (Certificado de Proficiência em Língua Portuguesa para Estrangeiros or Certificate of Proficiency in Portuguese for Foreigners), developed by the Brazilian Ministry of Education, is an official exam that certifies the proficiency of Brazilian Portuguese as a foreign language by assigning candidates to one of the four levels of proficiency: intermediate, upper intermediate, advanced or highly advanced. For detailed information, please consult: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/celpe-bras#

The LextPT test was administrated to a different group of participants for validation. The recruitment was made mainly through social media. We believe that such a less controlled form of subject recruitment leads to a relatively heterogeneous group of participants that can be more representative of the Portuguese L2 learners' diverse profiles. In total, 364 responses of Portuguese L1 speakers were collected in the validation study, 67 of which were later excluded for various reasons (48 bilinguals, four participated in the pilot study, five did not give consent, and ten completed the validation test more than once). The responses of 297 native participants were considered for analysis. This L1 group comprised 134 natives of EP (99 women, 33 men, 2 others; mean age = 33.6 years; range: 19.08–56.25, SD = 10.26) and 163 natives of BP (102 women, 60 men, 1 other; mean age = 34.29 years; range: 16.75–66.6, SD = 10.74). Of all L1 participants, 259 reported speaking at least one other language besides Portuguese (127 EP and 132 BP).

Five hundred and eight responses from L2 speakers of Portuguese were gathered in the validation phase. We only selected late L2 learners of BP or EP whose responses were considered valid. Before calculating the test scores, we removed 56 responses, because 12 participants did not give consent, 13 had participated in the pilot study, eight finished the test more than once, two were acquiring other varieties of Portuguese, 11 reported having studied Portuguese before adulthood, and ten selected all 90 items[11]. In total, we analysed the responses of 452 L2 participants, 270 learners of EP (202 women, 63 men, 5 others; mean age = 33.87 years; range: 18–71.22, SD = 15) and 182 learners of BP (113 women, 64 men, 5 others; mean age = 37.14 years; range: 18–70.8, SD = 12). The L1s of these participants are listed in Table 5.

The calculated LextPT scores of all of these 749 L1 and L2 participants are summarised in Table 6. The participants' performance in the validation study was in general similar to that in the pilot study. The mean accuracy of L1 speakers was significantly higher than that of L2 learners [Welch-corrected two-sample $t(667.59) = 32.851$, $p < .001$], with a large effect size [Cohen's $d = 2.14$]. These results are consistent with the between-group difference and effect size observed in the pilot phase, suggesting that the 90 items included in LextPT can effectively discriminate between L1 and L2 Portuguese speakers.

As in the pilot study, the L2 participants' test scores were first correlated against their self-assessment scores, visualised in Fig. 4. A moderate, but significant, correlation was again attested for learners of EP [Pearson's $r(268) = .5$, $p < .001$] as well as for learners of BP [Pearson's $r(180) = .44$, $p < .001$]. In addition, the L2 participants' LextPT scores were also moderately correlated with the years that they reported having spent on learning Portuguese (L2 EP group [Pearson's $r(264) = .42$, $p < .001$] and L2 BP group [Pearson's $r(176) = .35$, $p < .001$]). This correlation is illustrated in Fig. 5.

The L2 speakers' test scores were then plotted against the levels of their obtained Portuguese proficiency certificates or the levels of the Portuguese course they were taking. The data on certificate levels and on course levels were aggregated in the following way: Taking the CEFR criterion as an example, if a participant reported that she was enrolled in a B2 level course, she was considered to have the same proficiency as the one that had attained a B1 level certificate.

Among the 270 learners of EP, 149 provided information on their CAPLE exam levels (A1: 17; A2: 25; B1:28; B2: 37, C1:31; C2: 11). In the L2 BP group, 59 of the 182 participants indicated their level of CELPE-Bras (intermediate:

Table 5 Validation study: L2 group participants' native languages

L1	Number of participants	
	L2 EP	L2 BP
Spanish	13	107
Italian	13	9
French	20	5
Chinese	63	8
English	59	21
Croatian	23	1
Bulgarian	9	-
German	1	2
Dutch	1	1
Slovenian	1	2
Russian	3	4
Czech	3	-
Polish	27	4
Romanian	2	1
Swedish	2	-
Slovak	3	1
Galician	3	-
Finnish	1	2
Lithuanian	-	2
Turkish	2	-
Bilingual Catalan & Spanish	2	-
Bilingual Creole & French	2	-
NA	3	-

Others (one speaker of each)

L2EP: Afrikaans, Danish, Estonian, Malayalam, Norwegian, Ukrainian, Urdu, Wolof, Albanian & French, Croatian & Russian, Croatian & German, Polish & Russian, Ukrainian & Russian, Vietnamese & French & Lao.

L2BP: Arabic, Bengali, Cebuano, Haitian Creole, Indonesian, Konkani, Macedonian, Saamaka, English & Filipino, Italian & Russian, Spanish & Italian, Spanish & Guarani

[11] In line with Amenta et al. (2020), we consider that selecting all items instantiates a response strategy rather than a real performance on the test.

Table 6 Summary of the LextPT results of the validation study

	L1 EP	L1 BP	Overall L1	L2 EP	L2 BP	Overall L2
Mean	52.13	53.67	52.98	27.51	24.41	26.26
SD	7.55	6.18	6.86	14.36	15.94	15.08
Range	13–60	6–60	6–60	0–60	−25 to 60	−25 to 60

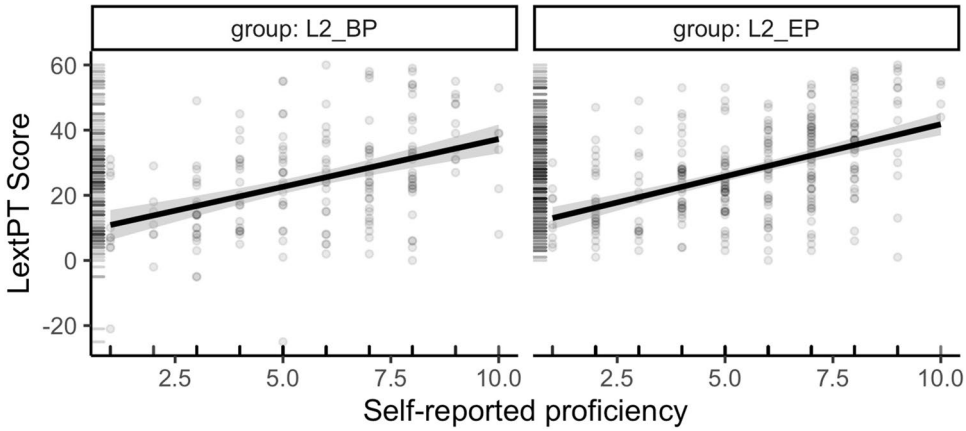

Fig. 4 LextPT scores (corrected accuracy; maximum 60) by self-rated Portuguese proficiency (1–10 scale; maximum 10) in the validation study

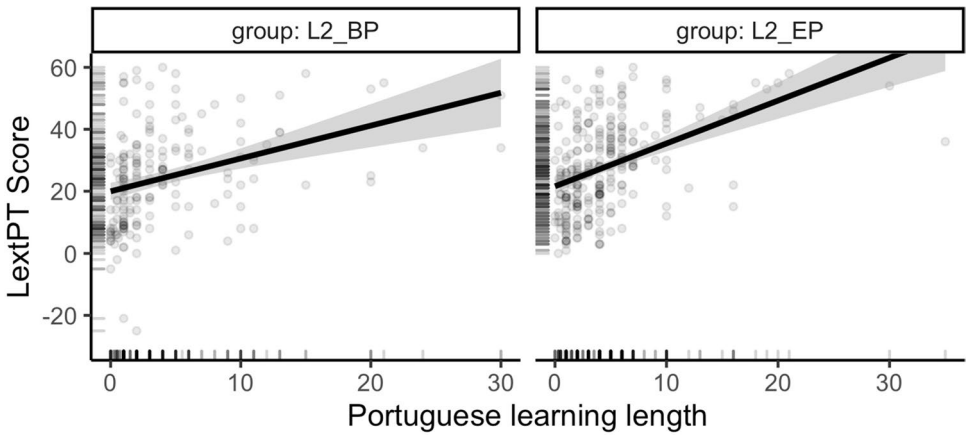

Fig. 5 LextPT scores (corrected accuracy; maximum 60) by Portuguese learning length (years) in the validation study

25; upper intermediate: 13; advanced: 21)[12]. Figures 6 and 7 respectively show that the performance of both the L2 EP and L2 BP groups was largely consistent with their proficiency levels. The same tendency was attested for the Italian test LexITA (Amenta et al. 2020).

Interestingly, similar to the validation results reported by Amenta et al. (2020), the A2-level participants behaved "unexpectedly" in comparison with participants from other levels. Amenta et al. (2020) reasoned that it might be the case that some participants used different criteria from their proficiency level to decide to which level they belonged, which, according to the authors, could be influenced by language anxiety. In our study, we consider another possibility. In the questionnaire, we asked the L2 participants to indicate the level of their obtained certificate, which makes it possible that the reported level of this certificate does not correspond to their current Portuguese

[12] One of the L2 learners of BP reported having the "highly advanced" certificate, but this was not sufficient for between-group comparison.

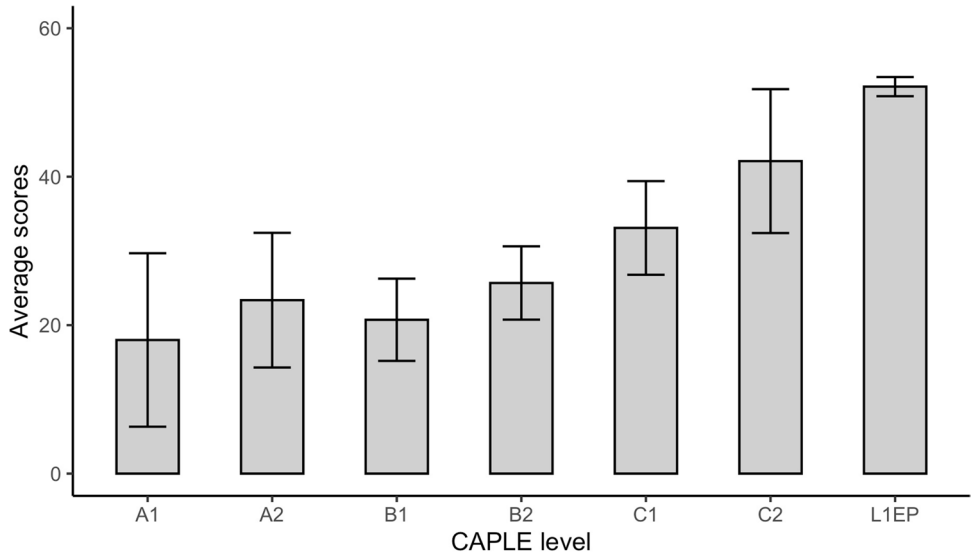

Fig. 6 Distribution of LextPT scores over CAPLE proficiency level. *Note.* Error bars show 95% confidence interval

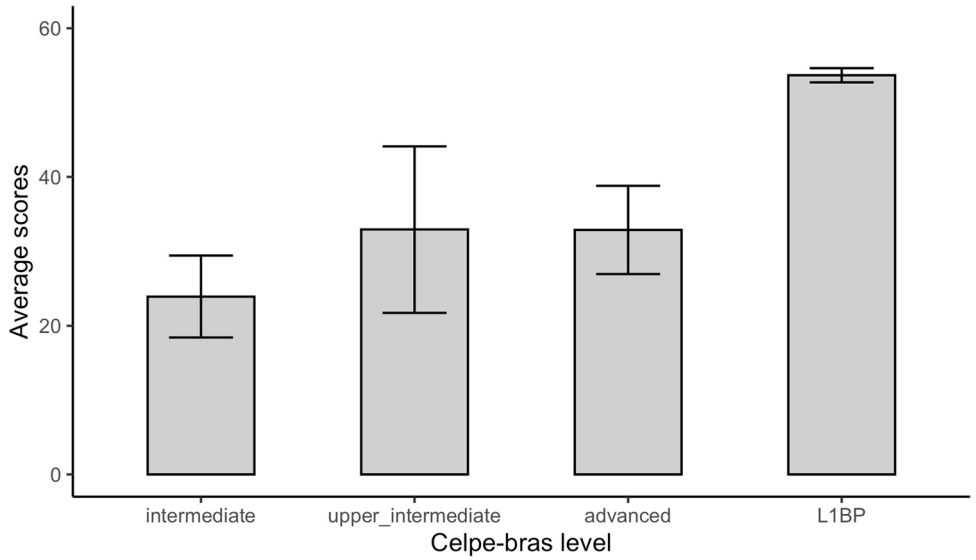

Fig. 7 Distribution of LextPT scores over CELPE-Bras proficiency levels. *Note.* Error bars show 95% confidence interval

proficiency because they might have obtained an elementary-level certificate several years ago and did not take a higher-level exam afterwards. For instance, one participant in the L2 EP group rated her Portuguese proficiency at 8 and reported having studied Portuguese for six years. Despite scoring 51 of 60 in LextPT, she indicated that she only had an A2 level certificate. Similar cases were observed in the participants who reported their CELPE-Bras proficiency levels. For example, one participant who reported an upper-intermediate level scored 58 of 60 in LextPT. According to the questionnaire, she had learned Portuguese for 15 years and rated herself at 8. This leads us to speculate that the reported CELPE-Bras level did not correspond to her actual proficiency in Portuguese. Such cases might explain the position of the mean scores between the upper-intermediate learners and the advanced learners.

Finally, the reliability of the test was assessed by computing Cronbach's alpha and the ICC coefficient. Taking both L1 and L2 participants into consideration, the test is overall very reliable [$\alpha = .96$; ICC3k = .96, $p < 0.001$]. The high reliability for both L2 EP group [$\alpha = .93$; ICC3k = .93, $p < 0.001$] and L2 BP group [$\alpha = .94$; ICC3k = .94, $p < 0.001$] suggests that LextPT is felicitous for reflecting the L2 Portuguese proficiency of both European and Brazilian varieties. The reliability analysis was further performed on the responses of learners of both varieties with typologically distinct L1s, namely Chinese (71 speakers), Germanic languages (90 speakers), Romance languages (175

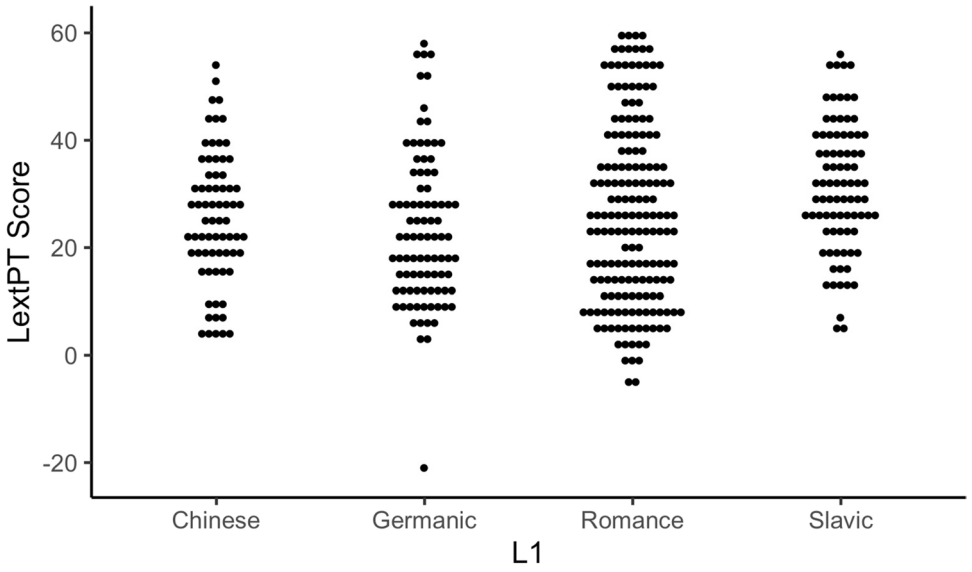

Fig. 8 The LextPT score distribution within each L1 group by typology

speakers), and Slavic languages (84 speakers). High reliability values were obtained for all three groups: Chinese [$\alpha = .87$; ICC3k = .87, $p < 0.001$], Germanic [$\alpha = .93$; ICC3k = .93, $p < 0.001$], Romance [$\alpha = .95$; ICC3k = .95, $p < 0.001$], and Slavic [$\alpha = .91$; ICC3k = .91, $p < 0.001$]. Moreover, LextPT test scores did not reach neither floor nor ceiling in any L1 groups by typology, but rather, are dispersed within each group, as shown in Fig. 8. These results revealed that LextPT is capable of discriminating the proficiency among learners with L1s that are either tightly related to Portuguese or not. All told as a whole, the final version of the LextPT test is sound for learners of both European and Brazilian varieties, with diverse linguistic backgrounds.

Discussion

As one of the most spoken languages in the world, Portuguese has seen a growing interest in learning it as a foreign language (Bateman et al., 2014; Sollai et al., 2018). Accordingly, an increasing number of studies have been implemented to better understand the L2 acquisition and processing of Portuguese in diverse contexts and by learners from different backgrounds (e.g. see Molsing et al., 2020 for a showcase). However, an efficient and reliable tool for measuring L2 Portuguese proficiency, especially in an experimental setting, has been lacking. Drawing inspiration from LexTALE (Lemhöfer & Broersma, 2012) and its equivalents (Brysbaert, 2013; Izura et al., 2014; Chan & Chang, 2018; Amenta et al., 2020), we reckon that it is timely to expand on previous studies and introduce LextPT, a quick and reliable vocabulary size test for the objective assessment of L2 Portuguese lexical knowledge, a good indicator of overall L2 proficiency in Portuguese, applicable to a wide range of learners in terms of both the varieties they acquire and their linguistic backgrounds.

Building on previous research, we adopted a careful item selection and evaluation paradigm:

(i) Word items were extracted in the criterion of frequency from subtitle-based lexical databases which closely resemble the daily usage of lexical items by native speakers; nonword stimuli were carefully selected in a way that they resemble the Portuguese ortho-phonotactic structure and, at the same time, do not have misleadingly great proximity to the Portuguese lexicon.
(ii) The 180 candidate items were first evaluated among a group of 250 L1 and L2 speakers (L1: 130; L2: 120) and only the most suitable 90 items (60 words and 30 nonwords) were included in the final version of LextPT, based on the point-biserial analysis and the IRT analysis on participants' responses.
(iii) The quality of the selected 90 items was further assessed in a validation study where the responses of 749 L1 and L2 speakers of Portuguese (L1: 297; L2: 452) were analysed.

Both the pilot and the validation results showed that the reliability of LextPT is quite high for L2 speakers. The comparative analyses between LextPT and other proficiency indicators, such as self-assessment scores, years of learning Portuguese as an L2, and existing proficiency test classification, further demonstrated that lexical knowledge may greatly vary within groups determined on the basis of other proficiency assessment

methods. LextPT thus displays a clear advantage in capturing the individual differences, allowing a more fine-grained distinction among learners. No ceiling effect nor floor effect was found for L2 speakers, suggesting that LextPT can assess a wide range of L2 Portuguese proficiency.

Although the development of LextPT strictly followed the standard procedure for extending the LexTALE paradigm to other languages, we expanded on previous studies in the following two aspects. First, we took both European and Brazilian varieties of Portuguese into consideration during the item selection and validation processes. This guarantees that LextPT can be employed to assess a larger group of L2 Portuguese learners. Second, responding to the calls in Izura et al. (2014) and Amenta et al. (2020) for collecting more norms for populations with different linguistic experience, we obtained a large-scale participant pool, which facilitated the investigation into whether LextPT is suitable for learners from different L1 backgrounds. We show that the high reliability of LextPT was consistent across different learner groups, such as native speakers of Romance languages (e.g. French, Italian, and Spanish), Germanic languages (e.g. Dutch, English, and German), Slavic languages (e.g. Bulgarian, Croatian, and Polish), and Chinese. Furthermore, the LextPT scores are dispersed in a comparable manner across learner groups, further highlighting the potential that L1-L2 similarity does not hinder the effectiveness of LexTALE tests.

We deem that, apart from providing information on participants' proficiency in Portuguese, the LextPT score can help to detect certain irregular participation. For example, in the validation study, one native EP speaker scored 13, one native BP speaker scored 6, and two L2 speakers respectively scored −21 and −25 points, which should raise researchers' attention to such anomalous performance in comparison with their peers.

In addition, like other versions of LexTALE, LextPT can be conducted in a few minutes. The rapid and easy implementation of LextPT allows it to be easily integrated into different kinds of experimental studies.

However, it is not the goal of this study to construct a test allowing horizontal comparisons of learners of different L1s, hence other factors were not controlled (e.g. age, language learning method). A conclusion cannot be drawn about to what extent the absolute scores obtained by bilinguals with different linguistic experience can be comparable. Further validation of LextPT can be conducted by correlating the LextPT scores against other well-established measurement of L2 Portuguese proficiency, or by looking into the predictive power of LextPT scores in experimental studies.

The creation of LextPT contributes to the effort to provide a rapid and reliable tool for the objective assessment of L2 proficiency across languages in the context of psycholinguistic research, initiated by Lemhöfer and Broersma (2012) and extended by Brysbaert (2013), Izura et al. (2014), Chan and Chang (2018), and Amenta et al. (2020). Although it is not the intention of the present study to replace any existing comprehensive proficiency test targeting different linguistic competences for different purposes, we believe that our effort brings forward the possibility of a more achievable comparison of results of experimental works across disciplines pertaining to Portuguese as an L2 via a standardised, efficient, and valid assessment tool.

Availability

LextPT is an effective and reliable assessment tool of L2 Portuguese vocabulary size (an important indicator of L2 proficiency). Its rapid (within 5 minutes) and flexible administration (either in electronic or paper format) facilitates easy integration into any experimental study. In line with previous LexTALE tests, we showed that presenting the test items of LextPT in either a random or fixed order does not influence its validity and reliability. This should alleviate the concern of some researchers who want to use the pen-and-paper version of LextPT and apply it in a fixed order to all participants.

A pen-and-paper version of LextPT, together with the instructions in either English or Portuguese, can be found in the Appendix to this paper.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-021-01731-1.

Acknowledgements We would like to thank João Veríssimo for his comments and suggestions on earlier versions of this article.

References

Amenta, S., Badan, L., & Brysbaert, M. (2020). LexITA: A Quick and Reliable Assessment Tool for Italian L2 Receptive Vocabulary Size. *Applied Linguistics,* 2020, 1-24. https://doi.org/10.1093/applin/amaa020

Astika, G. G. (1993). Analytical assessment of foreign students' writing. *RELC Journal*, 24, 61-72. https://doi.org/10.1177/003368829302400104

Barreiro, A., Wittmann, L. H., & Pereira, M. J. (1996). Lexical differences between European and Brazilian Portuguese. *INESC Journal of Research and Development*, *5*(2), 75–101.

Bateman, B. E., & de Almeida Oliveira, D. (2014). Students' Motivations for Choosing (or Not) to Study Portuguese: A Survey of Beginning-level University Classes. *Hispania*, *97*(2), 264–280. http://www.jstor.org/stable/24368776

Beglar, D., & Hunt, A. (1999). Revising and Validating the 2000 Word Level and University Word Level Vocabulary Tests. *Language Testing* 16(2). 131-162. https://doi.org/10.1177/026553229901600202

Bonk, W. (2000). Second Language Lexical Knowledge and Listening Comprehension. *International Journal of Listening,* 14(1), 14-31. https://doi.org/10.1080/10904018.2000.10499033

Brysbaert, M. & New, B. (2009). Moving beyond Kucera and Francis: A Critical Evaluation of Current Word Frequency Norms and the Introduction of a New and Improved Word Frequency Measure for American English. *Behavior Research Methods*, 41, 997-990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M. (2013). Lextale_FR a fast, free, and efficient test to measure language proficiency in French. *Psychologica Belgica*, 53(1), 23–37. https://doi.org/10.5334/pb-53-1-23

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*(2), 717–723. https://doi.org/10.3758/s13428-016-0728-y

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

Comrie, B. (1989). *Language Universals and Linguistic Typology: Syntax and Morphology,* Second Edition. The University of Chicago Press.

Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency effect in first- and second-language word recognition: a lexical entrenchment account. *Quarterly Journal of Experimental Psychology*, 66(5), 843–863. https://doi.org/10.1080/17470218.2012.720994

Eyckmans, J. (2004). *Measuring Receptive Vocabulary Size.* LOT.

Hulstijn, J. H. (2010). Measuring second language proficiency. In Blom E., & Unsworth, S. (eds): *Experimental Methods in Language Acquisition Research* (pp. 185–200). : Benjamins.

Izura, C., Cuetos, F., and Brysbaert, M. (2014). Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicológica,* 35(1), 49–66.

Justi, F. R. R., Justi, C. N. G., & Roazzi, A. (2014). Efeitos da similaridade ortográfica das pseudopalavras no acesso lexical [Pseudowords' orthographic similarity effects on lexical access]. *Arquivos Brasileiros de Psicologia* 66(3). 133-147. http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1809-52672014000300011&lng=pt&tlng=pt

Khare, V., Verma, A., Kar, B., Srinivasan, N., & Brysbaert, M. (2013). Bilingualism and the increased attentional blink effect: Evidence that the difference between bilinguals and monolinguals generalizes to different levels of second language proficiency. *Psychological Research* 77(6), 728–73. https://doi.org/10.1007/s00426-012-0466-4

Laufer, B. (1992). How much lexis is necessary for reading comprehension. In Arnaud P. J. L., & Béjoint, H. (eds.): *Vocabulary and applied linguistics* (pp. 126-132). : Macmillan.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different. *Applied Linguistics*, 19, 255-271. https://doi.org/10.1093/applin/19.2.255

LeBlanc, R. & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly* 19(4), 673–687. https://doi.org/10.2307/3586670

Leclercq, P. & Edmonds, A. (2014). How to assess l2 proficiency? An overview of proficiency assessment research. In Leclercq, P., Edmonds, A., & Hilton, H. (eds): *Measuring L2 Proficiency: Perspectives from SLA* (pp.3–23). https://doi.org/10.21832/9781783092291-004

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: a quick and valid Lexical Test for Advanced Learners of English. *Behavior research methods*, 44(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q). *Journal of Speech, Language and Hearing Research* 50(4), 940–967.

Mattos e Silva, R. V. (2013). O Português do Brasil [Brazilian Portuguese]. In: E. B. P. Raposo, E. B. P., Nascimento, M. A. C., Mota, L., & Mendes, A. (eds.), Gramática do português (pp.145-154). Lisboa: Fundação Calouste Gulbenkian.

Meara, P. (1992). *EFL vocabulary tests.* Swansea: Centre for Applied Language Studies, University of Wales Swansea.

Meara, P. (1996). The dimensions of lexical competence. In: Brown, G., K. Malmkjaer, & Williams, J. (eds.), *Performance and Competence in Second Language Acquisition* (pp. 35–53). Cambridge University Press.

Meara, P. & Buxton, B. (1987). An alternative to multiple choice vocabulary tests, *Language Testing*, 4, 142–154. https://doi.org/10.1177/026553228700400202

Milton J., Wade, J., & Hopkins, H. (2010). Aural word recognition and oral competence in a foreign language. In Chacón-Beltrán, R., Abello-Contesse, C., & Torreblanca-López, M. (eds.): *Further insights into non-native vocabulary teaching and learning* (pp. 83-98). : Multilingual Matters.

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In Bardel, C., Lindquist, C., & Laufer, B. (eds.): *L2 Vocabulary Acquisition, Knowledge, and Use: New Perspectives on Assessment and corpus Analysis* (pp. 57-78). Amsterdam, the Netherlands: European Second Language Association.

Molsing, K. V., Lopes Perna, C. B., & Tramunt Ibaños, A. M. (2020). *Linguistic Approaches to Portuguese as an Additional Language*. Amsterdam. John Benjamins Publishing Company.

Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge University Press.

Nation, I. S. P. (2013). Vocabulary Size in a Second Language. In Chapelle, C. A. (eds), *The Encyclopedia of Applied Linguistics*. https://doi.org/10.1002/9781405198431.wbeal1270

Nizonkiza, D. (2011). The relationship between lexical competence, collocational competence, and second language proficiency. *English Text Construction* 4(1), 113-145. https://doi.org/10.1075/etc.4.1.06niz9

Oscarson, M. (1989). Self-assessment of language proficiency: Rationale and applications. *Language Testing,* 6(1), 1–13. https://doi.org/10.1177/026553228900600103

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge. *The Canadian Modern Language Review*, 56, 282-307. https://doi.org/10.3138/cmlr.56.2.282

R Core Team. (2020). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Revelle, W. (2021). Psych: Procedures for psychological, psychometric, and personality research. Evanston, Illinois: Northwestern University. R package version 2.1.9. https://CRAN.R-project.org/package=psych.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1–25. https://doi.org/10.18637/jss.v017.i05

Segura, L. (2013). Geografia da língua portuguesa [Geographical distribution of Portuguese speakers]. In: E. B. P. Raposo, E. B. P., Nascimento, M. A. C., Mota, L., & Mendes, A. (eds.), *Gramática do português* (pp.145-154). Lisboa: Fundação Calouste Gulbenkian.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36(2), 139-152. https://doi.org/10.1080/09571730802389975

Strong-Klause, D. (2000). Exploring the effectiveness of self-assessment strategies in ESL placement. In Ekbatani, G., & Pierson, H. (eds.), *Learner-directed assessment in ESL* (pp. 49–73). : Lawrence Erlbaum Associates, Inc.

Soares, A. P., Machado, J., Costa, A., Iriarte, Á., Simões, A., de Almeida, J. J., Comesaña, M., & Perea, M. (2015). On the advantages of word frequency and contextual diversity measures extracted from subtitles: The case of Portuguese. *Quarterly journal of experimental psychology*, 68(4), 680–696. https://doi.org/10.1080/17470218.2014.964271

Sollai, S., Alvim, R., Bianconi, C. & Parma, A. (2018). Portuguese is in! From less commonly taught to critical to world language. *Todas as Letras Revista de Língua e Literatura.* 20. https://doi.org/10.5935/1980-6914/letras.v20n1p105-121.

Tang, K. (2012). A 61 Million Word Corpus of Brazilian Portuguese Film Subtitles as a Resource for Linguistic Research. In *University College London, Working Papers in Linguistics*, 25, 208-214.

Thomas, M. (1994). Assessment of L2 Proficiency in Second Language Acquisition Research. *Language Learning*, 44(2), 307–336. https://doi.org/10.1111/j.1467-1770.1994.tb01104.x

Tremblay, A. (2011). Proficiency assessment standard in second language acquisition research "Clozing" the gap. *Studies in Second Language Acquisition*, 33(3), 339–372. https://doi.org/10.1017/S0272263111000015

Venâncio, R. (2018). *Geração de Pseudopalavras para Avaliação Linguística [Pseudo-word Generator for Linguistic Evaluation].* (Unpublished master's thesis), University of Coimbra, Coimbra, Portugal.

Wetzels, W. L. , Costa, J., & Menuzzi, S. (2016). *The handbook of portuguese linguistics*. Malden: WILEY Blackwell.

Wittmann, L., Pêgo, T., & Santos, D. (1995). Português do Brasil e de Portugal: alguns contrastes. In *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística*, 465–487. Lisboa: APL/Colibri.

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship Between Lexical Competence and Language Proficiency: Variable Sensitivity. *Studies in Second Language Acquisition*, 27(4), 567–595. https://doi.org/10.1017/S0272263105050254

Zimmerman, K. J. (2004). *The role of Vocabulary Size in Assessing Second Language Proficiency*. (Unpublished master's thesis), Brigham Young University, Utah, USA.

Open Practices Statement

None of the data or materials for the experiments reported here are available, and none of the experiments were preregistered.