LEXTALE_CH: A Quick, Character-Based Proficiency Test for Mandarin Chinese

I Lei Chan and Charles B. Chang

1. Introduction

Given the growing interest within linguistics in how bilinguals, and second language (L2) learners more generally, process their L2 (e.g., Talamas et al., 1999; Prior et al., 2007), there has long been a need among language researchers for reliable and valid measures of L2 proficiency. Much of the research on L2 processing within experimental psychology (e.g., Arêas Da Luz Fontes & Schwartz, 2010; Leonard et al., 2010) has relied on questionnaire tools to provide the primary, or sometimes sole, source of data on L2 proficiency. Such questionnaires typically measure proficiency by asking participants for a subjective self-assessment of their own language ability, operationalized in terms of an overall rating and/or ratings by skill (i.e., listening, speaking, reading, writing). However, there are a number of problems with subjective assessments that limit their utility in experimental research. For one, even within one well-defined population (e.g., adult learners of the same L1 background acquiring the same L2 under the same learning conditions), learners at the same proficiency level may perceive their language abilities differently, leading to variation in self-assessments that does not reflect actual proficiency differences. Different studies also use different rating scales, such that proficiency measures are often not directly comparable across studies. Furthermore, the evidence for the external validity of self-assessments as measures of proficiency is relatively weak (Lemmon & Goggin, 1989; Delgado et al., 1999).

Thus, despite the fact that questionnaire-based methods of proficiency measurement are relatively simple and easy to administer, there has continued to be a need for reliable and valid methods of measuring proficiency—in particular, methods which can evaluate L2 proficiency as quickly and easily as a questionnaire. In the case of L2 English, Lemhöfer and Broersma (2012) proposed an objective alternative to subjective self-assessments, a vocabulary-based test called LexTALE (Lexical Test for Advanced Learners of English). LexTALE was designed as a standard and efficient tool for evaluating the English proficiency of L2 learners within a short amount of time (5–10 minutes). The test requires participants to identify the lexical status (i.e., real

* I Lei Chan and Charles B. Chang, Boston University. Corresponding author: Charles B. Chang, cc@bu.edu. We would like to thank the BU Center for the Humanities for funding, the audience at BUCLD 42 for helpful comments, and all study participants.

 Proceedings of the 42nd annual Boston University Conference on Language Development, ed. Anne B. Bertolini and Maxwell J. Kaplan, 114-130. Somerville, MA: Cascadilla Press.

word or nonce word) of 60 items (40 real words and 20 nonce words) by responding 'yes' (real) or 'no' (nonce). The set of test items and the test procedure were adopted from a longer (240-item) unpublished proficiency test used to test learners of high proficiency (Meara, 1996). To account for response bias (i.e., the tendency of L2 learners to identify unknown words as real words), the set of test items was designed to include a considerable number of plausible nonce words, in a real-to-nonce ratio of 2:1. Final scores are scaled to 100 points, with 80–100 being categorized as 'advanced', 60–79 as 'upper intermediate', and 59 and below as 'lower intermediate'.

Crucially, Lemhöfer and Broersma (2012) demonstrated the validity of LexTALE as a proficiency test by showing that LexTALE scores correlated more closely with data from translation tests and commercially available proficiency tests than did subjective self-assessments. They also found that LexTALE differentiated effectively between Dutch and Korean learners of English. Additional evidence of LexTALE's validity was reported by Diependaele et al. (2013), who observed a word frequency effect in visual word recognition which was more pronounced in L2 speakers than in native (L1) speakers, in accordance with their LexTALE scores (see Yap et al., 2008 for similar results). Findings of Khare et al. (2013) provided further support for the effectiveness of LexTALE. In this study, Hindi-English bilinguals performed a focal task (an attentional blink task), and their English proficiency was measured in terms of both self-ratings and LexTALE scores. Results showed a significant correlation between the outcome measure and English proficiency only when proficiency was measured in terms of LexTALE scores.

Besides English, LexTALE has also been made available in versions for other languages, such as Dutch and German (see Lemhöfer and Broersma, n.d.). The availability of the test in additional languages has introduced the advantageous possibility of score comparison across languages, although the Dutch and German versions of the test have not yet been normed or validated. Following Lemhöfer and Broersma (2012), other researchers joined the effort of expanding LexTALE to a wider set of languages and created an equivalent test in French (Brysbaert, 2013) and in Spanish (Izura et al., 2014). Both studies started off with more items, Brysbaert with 120 and Izura et al. with 180 (half real words and half nonce words in each case). Given the larger set of items, the researchers took into consideration the possible demotivation that might be felt on the part of low-proficiency participants who know very few words; thus, instead of being asked to make a yes/no decision on all items individually (the method in Lemhöfer and Broersma), participants in these latter studies were shown the whole list of items and asked to indicate (a) which words they knew and (b) which words they believed were real words in the target language (French or Spanish).

Both Brysbaert (2013) and Izura et al. (2014) chose their real words based on word frequency as the crucial criterion, which also helped to control the range of difficulty (following the same logic as Lemhöfer & Broersma, 2012) by virtue of the fact that high-frequency words should be known to all users of the

target language while low-frequency words should be known only to some L1 speakers. To carefully select the real words, both studies used a database of word frequencies based on film subtitles and selected words according to their frequency of occurrence in this database, including more lower-frequency words than high-frequency words. Word frequencies were further broken down into six categories: (a) less than 1 occurrence per million words (pm), (b) 1–5 pm, (c) 5–10 pm, (d) 10–20 pm, (e) 20–100 pm, and (f) more than 100 pm. After piloting, the French version of LexTALE (LEXTALE_FR) ended up with 84 items (56 real words, 28 nonce words), and the Spanish version (LEXTALE_ESP) with 90 items (60 real words, 30 nonce words). In both cases, a larger set of items was used than in the original LexTALE to improve the quality of the test.

Thus, LexTALE tests are currently available for English, Dutch, and German (Lemhöfer & Broersma, 2012), for French (Brysbaert, 2013), and for Spanish (Izura et al., 2014), but not yet for Chinese. Considering the rapid growth in global popularity of learning Chinese (Zhang & Lin, 2017), there is a need for an equivalent test in Mandarin Chinese, the standard variety that is typically the target variety for L2 learners. Even though there are several existing Chinese proficiency tests—notably China's *Hanyu Shuiping Kaoshi* (HSK; Teng, 2017) and Taiwan's Test of Chinese for Speakers of Other Languages (TOCFL; Chang, 2017), as well as numerous tests used in the US, including the Oral Proficiency Interview (OPI) developed by the American Council on the Teaching of Foreign Languages (ACTFL), ACTFL's Writing Proficiency Test (WPT), the Advanced Placement (AP) Chinese Language and Culture Test, and the SAT II Chinese Subject Test (Liu, 2017)—these tests differ significantly in format from LexTALE, making it difficult to compare scores on these tests to LexTALE scores in other languages. Furthermore, these tests take much longer to administer than LexTALE (in some cases, upwards of 60 minutes or more), which presents a challenge for including any of these within a study consisting of many other tasks, especially when proficiency measurement is not actually the main concern of the study.

Our aim in the present study was to develop a test for measuring proficiency (in particular, character-based lexical knowledge) in Mandarin Chinese that is as analogous to the original LexTALE as possible. In our view, a Mandarin version of LexTALE is likely to benefit linguistic research in three main ways. First, at a basic level it will provide a free, fast, and effective method for roughly estimating Mandarin vocabulary size. The final test described in this article can be administered in either a web-based or paper format and only takes about five minutes to complete, making it particularly suitable for low-stakes assessment. Second, a Mandarin LexTALE will allow easier comparison of Mandarin proficiency measures across studies of Mandarin, as well as easier comparison of proficiency levels in Mandarin with proficiency levels in other languages within studies of multilinguals. Finally, the information about variation in Mandarin lexical knowledge gathered through a Mandarin LexTALE may allow researchers to better account for individual differences in language processing in the Mandarin-speaking population (cf. Diependaele et al., 2013).

In the rest of this paper, we provide a detailed description of the test we developed, called LEXTALE_CH (i.e., a Lexical Test for Advanced Learners of Chinese), which we later used in the third language perception study presented at BUCLD 42. To develop LEXTALE_CH, we replicated the studies of Brysbaert (2013) and Izura et al. (2014), which served as a model for how to extend LexTALE to another language. As in these studies, the final set of items in LEXTALE_CH was selected through a pilot study. This pilot study started off with a large pool of 180 items (90 lexical characters, 90 nonce characters), which were tested on a group of L1 Mandarin speakers and a group of L2 Mandarin learners. The 90 items selected for LEXTALE_CH (60 lexical, 30 nonce) were those that could discriminate among these participants effectively.

2. Methods

2.1. Materials

The development of LEXTALE_CH started off with 180 Chinese-like characters: 90 real Chinese characters (i.e., lexical items) and 90 nonce characters (i.e., non-words). The 90 lexical items were selected through the method used by Izura et al. (2014), which was based on word frequency. The distribution of the lexical items across frequency tiers (in terms of occurrences per million characters; pm) is shown in Table 1.

Table 1: Distribution of real characters (lexical items) across frequency tiers.

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

The 90 lexical items were drawn from an online database compiled by Da (2004). This freely available database contains 193 million modern Chinese characters and is ordered according to character frequency (in pm). A list of 90 characters was compiled from this database, including both low-frequency characters and high-frequency characters. As seen in Table 1, the selected characters were skewed toward low-frequency tiers to increase the difficulty of the item set, because low-frequency characters should be harder to recognize than high-frequency ones for all Mandarin speakers, including L1 speakers.

The 90 nonce items were adopted from Peng et al. (1997) and varied in their orthographic plausibility. This is because in Chinese orthography, the position or arrangement of radicals in a character is non-arbitrary. For example, in horizontally oriented characters, some radicals appear only on the left side (left-position radicals) whereas other radicals appear only on the right side (right-

position radicals). Thus, there were two types of nonce items: 45 'pseudo-characters' fully complied with Chinese orthographic rules (i.e., with radicals appearing in the correct position), while 45 'non-characters' violated one or more of these rules (i.e., with radicals appearing in an incorrect position). Our expectation was that pseudo-characters, since they are plausible characters from the point of view of compliance with orthographic conventions, would be relatively more difficult to identify correctly as nonce items.

2.2. Procedure

The pilot study consisted of a web-based test that participants took at a convenient time and place. This method allowed for the collection of more norming data for the selection of final test items. The pilot test was administered using Qualtrics Survey Software (Qualtrics, 2016). The items were arranged into a random order for test presentation and shown in that order to all participants.

Prior to beginning the pilot test, participants were given detailed written instructions in both Simplified Chinese and English regarding the lexical decision task (see Appendix A). In particular, they were told to indicate which Chinese characters they knew and, additionally, which characters they believed to be a real Chinese character (regardless of knowing the character or not).

As shown in Appendix A, pilot participants were informed that they were going to see a series of Simplified Chinese characters, some of which corresponded to real Chinese words and some of which did not. During the test, a set of 180 Simplified Chinese characters was presented on screen, and participants had to check a box above the character if they identified it as an authentic Chinese character. Crucially, participants were asked explicitly to complete the test on their own (i.e., without the aid of a dictionary) as the data would otherwise not be informative. They were also told that they had as much as time as they needed to complete the test (i.e., the task was not speeded).

After the test, participants were asked to fill out a background questionnaire, which elicited information about their gender, native language, length of time using/learning Mandarin, self-rated overall Mandarin proficiency (on a 10-point scale), and self-rated Mandarin proficiency by language skill (i.e., listening, speaking, reading, writing; each on a 10-point scale).

Following the pilot study, which determined the set of items included in LEXTALE_CH, a validation study was carried out to check the quality of the final items in the absence of the excluded items. The format and administration of the validation test was identical to the pilot test, except that only the 90 items selected for LEXTALE_CH were shown to participants.

2.3. Participants

Because the pilot test measured vocabulary size using Simplified Chinese characters, the sample of participants included in the norming analyses was limited to Mandarin speakers who had acquired Simplified Chinese. Thus, L1

Mandarin speakers in this sample were those born and educated in mainland China (where the writing system of Simplified Chinese characters is officially used); participants educated in Hong Kong, Macau, or Taiwan were excluded since the writing system of Traditional Chinese characters is officially used in these regions. Similarly, the L2 Mandarin learners in this sample were those who had learned Simplified Chinese.

A total of 310 pilot tests were started, of which 64 were completed by participants meeting the requirements of the study. Participants were first filtered by a prescreening questionnaire, which classified them into three groups: (a) L1 Mandarin speakers raised in mainland China, (b) L1 Mandarin speakers raised in Hong Kong, Macau, and/or Taiwan, and (c) L2 Mandarin speakers. Only groups (a) and (c) were allowed into this study; group (b) was excluded. Thus, the final sample of pilot participants consisted of a total of 64 self-identified adult Mandarin speakers: 49 L1 Mandarin speakers and 15 L2 Mandarin learners. According to the post-test questionnaire data, the L2 Mandarin learners came from various L1 backgrounds: English (N = 9), Danish (N = 1), German (N = 1), Indonesian (N = 1), Mauritian Creole (N = 1), Nepali (N = 1), and Spanish (N = 1). The L2 participants tended to be highly experienced learners of Mandarin, reporting several years of study on average, although there was also considerable variation in their length of study (M = 5.0 yr, SD = 7.2). On average, they rated their Mandarin proficiency (on a 10-point scale) as intermediate, although again there was substantial variation in self-rated proficiency (M = 5.3, SD = 2.0).

As for the validation study, a total of 114 validation tests were started, of which 94 were completed by participants meeting the requirements of the study. Thus, the final sample of validation participants consisted of 94 self-identified adult Mandarin speakers: 69 L1 Mandarin speakers and 25 L2 Mandarin learners. None of the participants in the validation study had participated in the initial pilot study. Post-test questionnaire data indicated that the L2 Mandarin learners again came from various L1 backgrounds: English (N = 16), German (N = 2), Hindi (N = 1), Indonesian (N = 1), Korean (N = 1), Russian (N = 1), Spanish (N = 1), Tagalog (N = 1), and Vietnamese (N = 1). These L2 participants also tended to be highly experienced learners of Mandarin; like the L2 participants in the pilot study, they reported several years of study on average, with considerable variation in their length of study (M = 5.7 yr, SD = 3.8). On average, they rated their Mandarin proficiency (on a 10-point scale) as intermediate as well (M = 5.5, SD = 1.7).

3. Results

The quality of the test items was assessed in two stages. First, the relationship of each item to a participant's overall responses to the items was examined using point-biserial correlation and Item Response Theory (IRT) analysis. Second, the item was examined in the context of the larger participant sample: Cronbach's alpha was used to measure item reliability by viewing

participants' responses as a group, and criterion validity was measured by comparing the performance of L1 and L2 Mandarin speakers. Each of these sets of analyses is discussed in turn below.

3.1. Initial item assessment

The data collected in the pilot study were analyzed using the same methodology as Brysbaert (2013) and Izura et al. (2014), which examined the quality of each item using point-biserial correlation and IRT analysis. First, using the *ltm* package (Rizopoulos, 2006) in R (R Development Core Team, 2013), each item's predictiveness (with respect to overall test score) was assessed by calculating the point-biserial correlation between participants' responses to the item and their overall scores. This correlation ranged between -1 and +1: a positive correlation indicated that a high performer (i.e., a participant who obtained a high overall test score) tended to perform better on the given item than a low performer, while a negative correlation revealed an anomalous situation in which a high performer tended to perform less well on the given item than a low performer. Out of the 180 piloted items, 21 showed a negative point-biserial correlation, meaning that they were more likely to be identified as a lexical item by high performers (i.e., participants with good knowledge of Chinese) than low performers. These 21 items comprised 18 nonce items (namely, pseudo-characters) and 3 lexical items of very low frequency (less than 1 pm). Because these items did not predict overall score in the desired manner, they were removed from subsequent analyses.

The remaining 159 items then underwent IRT analysis (also using the *ltm* package in R), which provided information about each item's difficulty and discrimination power. Discrimination power refers to how well an item can distinguish a high performer from a low performer. Thus, IRT analysis takes an individual's performance as well as the difficulty of the item into account. Based on the IRT analysis, 60 lexical items and 30 nonce items of various difficulty levels, all with good discrimination power, were selected (see Appendix B for the full set of selected items). The selection was done by ordering the items according to difficulty level, dividing them into equal intervals (i.e., groups containing an approximately equivalent number of items each), and then selecting the item with the best discrimination power from each interval. The 90 items selected in this manner were the basis for participants' overall test scores (discussed in the next section).

3.2. Reliability and validity

In accordance with Lemhöfer and Broersma (2012), Brysbaert (2013), and Izura et al. (2014), there are two ways to calculate a participant's overall score on LEXTALE_CH: (1) Raw Accuracy and (2) Corrected Accuracy.

Raw Accuracy is a mean proportion of correct responses. It is calculated according to the formula in (1) below. Because the total number of nonce items

in LEXTALE_CH (30) is only half of the total number of lexical items (60), a participant's number of correct responses on nonce items (*b*) is multiplied by 2 in order to weight performance on lexical items and nonce items equally; thus, the division in (1) is by 120 instead of by 90. As a proportion measure, Raw Accuracy ranges from 0 to 1, with a score lower than .5 indicating that the participant was more likely to respond incorrectly than correctly.

(1) Raw Accuracy = $(a + 2 * b) / 120$
where a = number of correct responses for lexical items
where b = number of correct responses for nonce items

The second calculation, Corrected Accuracy, has the advantage of penalizing guessing behavior. It is calculated as in (2) below, where, as in (1), the nonce item component is multiplied by a factor of 2 to account for the lower number of nonce items. If participants respond without regard to the items (e.g., completely randomly, or saying 'yes' indiscriminately), their Corrected Accuracy is expected to come out to around 0 (as opposed to around .5 for Raw Accuracy). On the other hand, if participants (incorrectly) accept nonce items as real Chinese at a higher rate than they do lexical items, their Corrected Accuracy is expected to be negative, as low as -30. According to (2), the highest possible Corrected Accuracy score is 60, which requires participants to identify all lexical items correctly (i.e., get all 'hits') and not to select any nonce items incorrectly (i.e., avoid any 'false alarms'). Thus, most participants who perform the lexical decision task in good faith are expected to obtain scores somewhere between 0 and 60. To provide an example of intermediate performance, a participant who correctly selects 55 lexical items and incorrectly selects 7 nonce items would obtain a Corrected Accuracy score of 41 (= 55 – 2 * 7).

(2) Corrected Accuracy = $h – 2 * f$
where h = number of correctly identified lexical items ('hits')
where f = number of incorrectly accepted nonce items ('false alarms')

Using the formulae in (1) and (2), Raw Accuracy and Corrected Accuracy scores were calculated for all 64 pilot participants, and the results are summarized in Table 2. Corrected Accuracy in this sample ranged from 0 to 55.

Table 2. Summary of LEXTALE_CH results from the pilot study.

Variable (possible range)	L1 Mandarin speakers		L2 Mandarin speakers	
	Mean (SD)	Range	Mean (SD)	Range
Raw accuracy (0–1)	.87 (.05)	.74–.96	.63 (.10)	.50–.82
Hit count (0–60)	54.9 (5.2)	33–60	29.1 (9.6)	17–43
False alarm count (0–30)	5.5 (3.5)	0–13	6.8 (4.0)	2–15
Corrected accuracy (-30–60)	43.9 (5.9)	29–55	15.5 (11.9)	0–38

A large, and statistically significant, difference was observed between the L1 and L2 groups [M_{L1} = 43.9, M_{L2} = 15.5; Welch-corrected two-sample $t(16.1)$ = 8.902, $p < .001$], with an effect size (Cohen's d; Cohen, 1988) of 2.91. This difference accords with the differences found in Brysbaert (2013) for LEXTALE_FR and in Izura et al. (2014) for LEXTALE_ESP.

To examine the relationship between LEXTALE_CH results and self-assessments of proficiency, participants' LEXTALE_CH scores (in terms of Corrected Accuracy) were correlated against their Mandarin proficiency self-assessments (specifically, a rating of overall Mandarin proficiency on a 1–10 scale). This analysis showed a significant, and large, global correlation between LEXTALE_CH scores and proficiency self-ratings [Pearson's $r(62) = .78$, $p < .001$], due in large part to the pronounced differences between the L1 and L2 groups in both dimensions. The correlation was smaller within the L1 group [Pearson's $r(47) = .30$, $p < .05$] than within the L2 group [Pearson's $r(13) = .63$, $p < .05$], as shown in Figure 1. Correlations of LEXTALE_CH scores against self-ratings of Mandarin reading proficiency specifically were very similar.

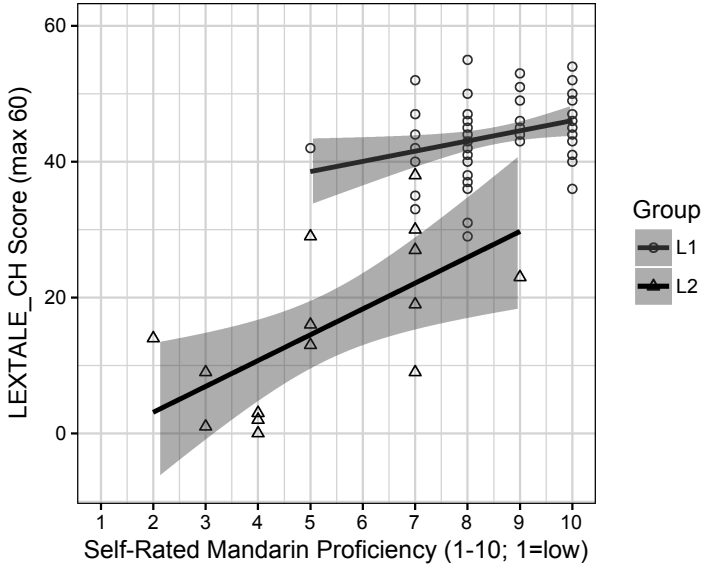


Figure 1: LEXTALE_CH scores (Corrected Accuracy; maximum 60) by self-rated Mandarin proficiency (1–10 scale; maximum 10) in the pilot study. L1 speakers are shown in circles; L2 speakers, in triangles. Each shaded area represents the 95% confidence interval around the regression line.

L2 speakers who rated their overall Mandarin proficiency at lower than 5 obtained systematically lower LEXTALE_CH scores, suggesting they had weaker lexical knowledge of Mandarin, whereas L1 speakers who rated their

proficiency at 7–10 generally obtained high scores. Interestingly, L2 speakers who rated their proficiency at 6–7 scored similarly to L2 speakers who rated their proficiency at 5, and lower than L1 speakers who rated their proficiency at the same level. Similarly, in the L1 group, some participants who rated their proficiency at 10 (i.e., the highest possible level of Mandarin proficiency) scored no higher than the ones who rated their proficiency lower, at 7–9. Together, these results are consistent with the view that subjective self-assessments of proficiency are related to actual proficiency (as measured by objective performance assessments), but are not a perfect reflection of actual proficiency.

The last analysis performed on the data from the pilot study was a reliability analysis. Reliability (i.e., internal consistency) was measured in terms of Cronbach's alpha (α; Cronbach, 1951), which takes responses on all test items into consideration. Including both L1 and L2 participants in this analysis showed that LEXTALE_CH had excellent reliability [$\alpha = .95$]. Reliability was still high [$\alpha = .86$] when considering only L2 participants, although intermediate [$\alpha = .66$] when considering only L1 participants. The reason for the lower reliability of the test for L1 speakers is not clear; however, it may be due in part to educated adult L1 speakers showing less variability in character-based lexical knowledge than L2 learners (cf. lower *SD* of Corrected Accuracy for L1 than L2 speakers; Table 2). This type of disparity could make variation in scores on a test such as LEXTALE_CH end up reflecting, for L1 speakers, individual differences in other, non-focal variables (e.g., visual acuity, attention) to a greater degree.

In sum, the results of the pilot study suggested that the final 90 test items (selected through the item assessment described above) combine to make an effective test for measuring Mandarin proficiency, as indexed by character-based receptive vocabulary, in a short amount of time. However, recall that the pilot study presented this set of items within the context of a larger pool of 180 candidate items; as such, the results of the pilot study provided only suggestive evidence of the test items' effectiveness in the intended context (i.e., on their own, with no extra items). Therefore, in order to confirm that the final test set would work on its own, we conducted a validation study, using all and only the final 90 items and testing a different sample of L1 and L2 Mandarin speakers.

The results of the validation study are summarized in Table 3, which shows that performance in the validation study was overall similar to performance in the pilot study. L1 speakers attained a mean Corrected Accuracy (42.5) that was slightly lower than that of L1 speakers in the pilot study (43.9). On the other hand, L2 speakers' mean Corrected Accuracy (17.3) was slightly higher than that of L2 speakers in the pilot study (15.5), resulting from a higher number of both hits and false alarms. Nevertheless, the difference in performance between the L1 and L2 groups in the validation study was still significant [Welch-corrected two-sample $t(27.2) = 10.474$, $p < .001$], with a very large effect size [Cohen's $d = 2.75$]. These results suggest that the 90 test items comprising LEXTALE_CH were similarly effective in distinguishing between L1 and L2 speakers in the validation study as in the pilot study; that is to say, they were not affected by the absence of the items that were removed following the pilot study.

Table 3: Summary of LEXTALE_CH results from the validation study.

Variable (possible range)	L1 Mandarin speakers		L2 Mandarin speakers	
	Mean (SD)	Range	Mean (SD)	Range
Raw accuracy (0–1)	.85 (.04)	.75–.93	.64 (.10)	.44–.84
Hit count (0–60)	50.8 (6.3)	32–60	41.2 (10.2)	25–56
False alarm count (0–30)	4.1 (3.2)	0–13	11.9 (5.2)	3–27
Corrected accuracy (-30–60)	42.5 (4.9)	30–52	17.3 (11.6)	-7–41

As in the pilot study, participants' LEXTALE_CH scores in the validation study were correlated against their overall Mandarin proficiency self-assessments from the post-test questionnaire. This analysis, too, showed a large global correlation between LEXTALE_CH scores and proficiency self-ratings [Pearson's $r(92) = .67$, $p < .001$]. However, as in the pilot study, the correlation was weaker—in fact, not significant—within the L1 group [Pearson's $r(67) = -.10$, $p = .421$] than within the L2 group [Pearson's $r(23) = .43$, $p < .05$], as shown in Figure 2. Correlations of LEXTALE_CH scores against self-ratings of Mandarin reading proficiency specifically showed a similar pattern.

Figure 2: LEXTALE_CH scores (Corrected Accuracy; maximum 60) by self-rated Mandarin proficiency (1–10 scale; maximum 10) in the validation study. L1 speakers are shown in circles; L2 speakers, in triangles. Each shaded area represents the 95% confidence interval around the regression line.

The last analysis performed on the data from the validation study was, as in the pilot study, an analysis of reliability (i.e., internal consistency). This analysis

showed a very similar pattern of results to the pilot study: LEXTALE_CH had high reliability [α = .88] when considering both L1 and L2 participants, high reliability [α = .83] when considering only L2 participants, and intermediate reliability [α = .63] when considering only L1 participants.

4. Discussion

In this paper, we described the development and validation of a quick, character-based lexical test for objective assessment of proficiency in Mandarin Chinese, LEXTALE_CH. Inspired by LexTALE (Lemhöfer & Broersma, 2012), a similar test for objective assessment of English proficiency, LEXTALE_CH follows LexTALE adaptations for Dutch and German, as well as for French (LEXTALE_FR; Brysbaert, 2013) and Spanish (LEXTALE_ESP; Izura et al., 2014), in extending the LexTALE paradigm to Mandarin Chinese, a widely studied language for which an equivalent test does not yet exist. Development of LEXTALE_FR and LEXTALE_ESP involved careful selection and testing of stimulus materials and, in the case of LEXTALE_ESP, additional validation of the test materials as well. Consequently, we also undertook a multi-step procedure to develop the materials for LEXTALE_CH, which closely resembled the process of creating the original LexTALE, LEXTALE_FR, and LEXTALE_ESP: (a) initial, frequency-based selection of possible lexical test items, along with creation of possible nonce test items, (b) a pilot study testing L1 and L2 speakers of Mandarin on a large pool of 180 candidate items, and (c) a validation study testing L1 and L2 speakers of Mandarin on the final set of 90 test items only. For reasons of continuity with LEXTALE_FR and LEXTALE_ESP, we have named this test LEXTALE_CH (LEXical Test for Advanced LEarners of CHinese).

Although the overall format of LEXTALE_CH is very similar to that of other versions of LexTALE, it should be noted that LEXTALE_CH differs from these other versions in two respects. First, LEXTALE_CH contains a larger set of test items (90 as opposed to 60), resulting in a test of higher reliability (α = .95; cf. α = .81 for LexTALE). Nevertheless, the test takes a similar amount of time (five minutes on average in the validation study) and produces a similar range of scores (going up to a maximum Corrected Accuracy score of 60). Second, unlike previous versions of LexTALE, LEXTALE_CH contains two types of nonce items, pseudo-characters and non-characters, meaning that some of the nonce items are orthographically impossible (i.e., violate orthographic rules) in Chinese. This contrasts with other versions of LexTALE, which include only phonotactically and orthographically possible nonce items.

The reason for using only orthographically possible nonce items in a lexical decision task is to increase the difficulty of the task (since it is trickier to correctly rule out as a non-word an item which is completely plausible phonotactically and orthographically), and our reasoning behind considering non-characters (i.e., orthographically impossible items) for inclusion in LEXTALE_CH was complementary: to moderate the difficulty of the test.

Because Chinese is one of the most difficult languages for English speakers to acquire (Lett & O'Mara, 1990), and was the first (and in most cases only) non-alphabetic L2 to be learned by our participants more generally, we proceeded with the development of nonce items cautiously, by creating and testing both pseudo-characters and non-characters. In the end, the nonce items with the best discrimination power comprised 23 pseudo-characters and 7 non-characters. As expected, L1 Mandarin speakers, both in the pilot study and in the validation study, were found to make more errors on the pseudo-characters than on the non-characters. The same effect was found among L2 speakers in both studies.

This brings us to the principal distinguishing characteristic, and limitation, of LEXTALE_CH: the orthographic basis of the test in the quasi-logographic Chinese writing system. Clearly, it is possible for acquisition of Mandarin—or, for that matter, any target language—to occur without acquisition of literacy; however, LEXTALE_CH, like other versions of LexTALE, is a written test that relies on the participant to be literate in the target language. Therefore, it is important to note that LEXTALE_CH produces a measure of Mandarin proficiency in functionally literate speakers specifically and is not intended to be used with speakers who are illiterate or dyslexic. As such, the population of Mandarin speakers and learners targeted by this test is composed primarily of individuals who have significant formal educational exposure to written Chinese characters. The scope of the test thus inevitably excludes some non-trivial portion of the Mandarin-speaking population, such as L2 learners who received exposure to spoken Mandarin only or who are familiar with one or more transliteration systems but not with characters.

In connection with the written administration of LEXTALE_CH, the applicability of the test is also limited by the specific character set that we used—namely, Simplified Chinese. We opted to create LEXTALE_CH in Simplified Chinese over Traditional Chinese due to the greater numbers of Mandarin speakers worldwide who read and write primarily Simplified Chinese. However, as previously mentioned, this means that the current version of LEXTALE_CH will be inappropriate for Mandarin speakers who read and write primarily Traditional Chinese, such as speakers from Hong Kong, Macau, or Taiwan. It remains a project for future research to develop an additional version of LEXTALE_CH in Traditional Chinese.

5. Conclusion

The contribution of this study is in developing an objective, yet practical, performance-based assessment of proficiency in Mandarin Chinese, LEXTALE_CH. With over one billion speakers worldwide (Simons & Fennig, 2017), including a large and growing population of L2 speakers, Mandarin is a language for which a concise and accessible proficiency measurement tool such as LEXTALE_CH is long overdue. To our knowledge, LEXTALE_CH is the first proficiency testing instrument for Mandarin Chinese with good reliability and validity that can be completed in a matter of minutes. Our findings from

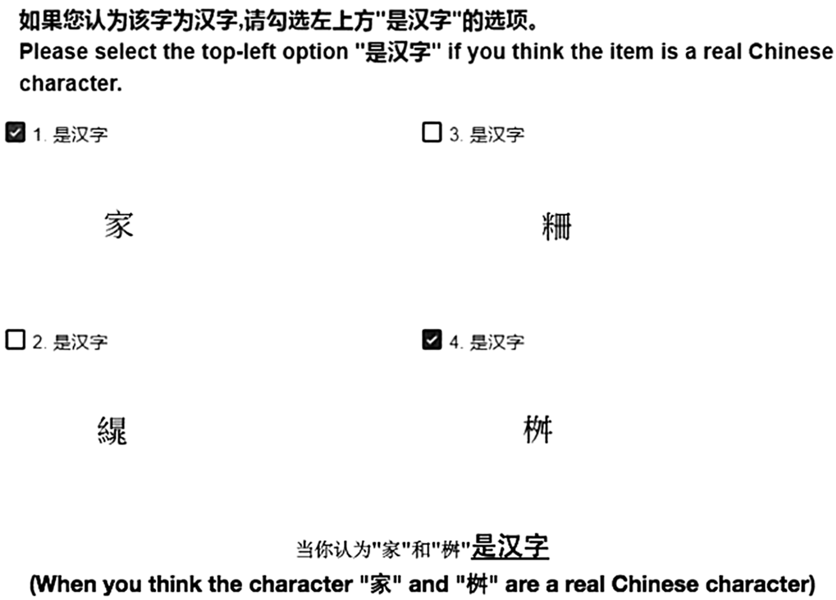
both piloting and validation indicate that LEXTALE_CH does not lead to ceiling effects among L1 speakers or floor effects among L2 learners, consistent with the wide span in item difficulty. As such, LEXTALE_CH has the potential to be a helpful tool for a diverse range of research projects needing to assess L1 and/or L2 Mandarin proficiency in a rapid and reliable manner.

In closing, we would like to emphasize that LEXTALE_CH is not intended to replace in-depth proficiency assessments, such as the various standardized Chinese tests mentioned in §1. On the contrary, when a given research question involves obtaining a nuanced picture of an individual's Mandarin proficiency, it will usually be more appropriate to use a longer, and multi-modality, Chinese test rather than a quick assessment of receptive vocabulary based on character recognition only. Our point is that when the research question does not require such a detailed picture of proficiency and/or the study cannot devote a great deal of time to proficiency assessment (for example, when proficiency is being measured as one of several control factors rather than as the main dependent variable), it is still possible, as well as practical, to obtain an objective, performance-based assessment of Mandarin proficiency, as opposed to one based solely on a subjective self-report. It is our hope that LEXTALE_CH will prove useful to the community of researchers in language acquisition, multilingualism, and Chinese linguistics and inspire further development of practical, performance-based assessments of proficiency in additional languages.

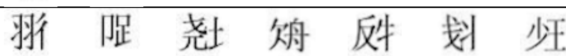
Appendix A: Instructions to participants in the LEXTALE_CH pilot study

在下一页，您将会看到 180 个看上去像"汉字"的字，当中只有一些是真正存在的汉字。您需要对每一个字做出判断，如果您认为该字是在中文里存在的(即使您不能够明确地说出该字的意思)或者是您知道该字的话，请勾选左上方的**"是汉字"**选项 。您无需快速回答每一道问题，但请您根据您的第一反应来作答，不用过度的犹豫。请在没有任何外来帮忙的情况下独立完成此测试（**不要使用任何汉语词典！**）。所有的字皆为**简体中文**。下方的图片为作答例子。

On the next page, you are going to be shown 180 characters that look "Chinese". Only some of them are real Chinese characters. You have to decide for each item whether it is a real Chinese character or not. Please select the top-left option "是汉字" if you know the character or if you believe that the item is a real Chinese character, even if you may not know its precise meaning. You do not need to respond rapidly, but please give your first impression, without any outside aids (i.e., **do not consult a dictionary!**). Each character will be presented in **Simplified Chinese**. A sample of what the test looks like is shown below.



Appendix B: Lexical items and nonce items in LEXTALE_CH

Lexical items (*N* = 60)									
淬	纑	舫	毸	麓	级	殛	褙	樵	鄣
篌	笵	篾	腋	跚	嗫	椰	颚	俪	乓
痹	聿	烹	辜	坤	劈	虏	匈	秃	悼
奉	滋	鸣	掠	恨	龄	吙	逖	堔	骾
阕	稻	镵	惝	瞭	凇	昏	欢	峒	蟑
踊	冗	坪	桨	隧	涕	隅	朔	夜	乖
Nonce items (*N* = 30)									
Pseudo-characters									
Non-characters									

Note: Instructions to the participant, along with the format of the web-based presentation of the items, are shown in Appendix A. Researchers who prefer to administer the test in a paper-based version may download a PDF file of the full test (with instructions in either English or Chinese), including an answer key, at https://osf.io/r3vs9/.

References

Arêas Da Luz Fontes, Ana B. & Schwartz, Ana I. (2010). On a different *plane*: Cross-language effects on the conceptual representations of within-language homonyms. *Language and Cognitive Processes, 25*(4), 508–532.

Brysbaert, Marc. (2013). LEXTALE_FR: A fast, free, and efficient test to measure language proficiency in French. *Psychologica Belgica, 53*(1), 23–37.

Chang, Li-ping. (2017). The development of the Test of Chinese as a Foreign Language (TOCFL). In Zhang & Lin (Eds.), pp. 21–42.

Cohen, Jacob. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, Lee J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–344.

Da, Jun. (2004). *Modern Chinese character frequency list* (现代汉语单字频率列表). https://github.com/thyrlian/namedict/blob/master/data/Modern%20Chinese%20Character%20Frequency%20List. Last retrieved November 15, 2016.

Delgado, Pedro, Guerrero, Gabriela, Goggin, Judith P., & Ellis, Barbara B. (1999). Self-assessment of linguistic skills by bilingual Hispanics. *Hispanic Journal of Behavioral Sciences, 21*(1), 31–46.

Diependaele, Kevin, Lemhöfer, Kristin, & Brysbaert, Marc. (2013). The word frequency effect in first- and second-language word recognition: A lexical entrenchment account. *The Quarterly Journal of Experimental Psychology, 66*(5), 843–863.

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

Khare, Vatsala, Verma, Ark, Kar, Bhoomika, Srinivasan, Narayanan, & Brysbaert, Marc. (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), 728–737.

Lemhöfer, Kristin & Broersma, Mirjam. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods, 44*(2), 325–343.

Lemhöfer, Kristin & Broersma, Mirjam. (n.d.). *LexTALE*. http://www.lextale.com/. Last retrieved August 23, 2017.

Lemmon, Christian R. & Goggin, Judith P. (1989). The measurement of bilingualism and its relationship to cognitive ability. *Applied Psycholinguistics, 10*(2), 133–155.

Leonard, Matthew K., Brown, Timothy T., Travis, Katherine E., Gharapetian, Lusineh, Hagler, Donald J., Jr., Dale, Anders M., Elman, Jeffrey L., & Halgren, Eric. (2010). Spatiotemporal dynamics of bilingual word processing. *NeuroImage, 49*(4), 3286–3294.

Lett, John A. & O'Mara, Francis E. (1990). Predictors of success in an intensive foreign language learning context: Correlates of language learning at the Defense Language Institute Foreign Language Center. In Thomas S. Parry & Charles W. Stansfield

(Eds.), *Language aptitude reconsidered*, pp. 222–260. Englewood Cliffs, NJ: Prentice Hall Regents.

Liu, Yan. (2017). Assessing Chinese in the USA: An overview of major tests. In Zhang & Lin (Eds.), pp. 43–65.

Meara, Paul M. (1996). *English vocabulary tests: 10k*. Unpublished manuscript. Swansea: Center for Applied Language Studies.

Peng, Dan-ling, Li, Yan-ping, & Yang, Hui. (1997). Orthographic processing in the identification of Chinese characters. In Hsuan-Chih Chen (Ed.), *Cognitive processing of Chinese and related Asian languages*, pp. 85–108. Hong Kong: The Chinese University Press.

Prior, Anat, MacWhinney, Brian, & Kroll, Judith F. (2007). Translation norms for English and Spanish: The role of lexical variables, word class, and L2 proficiency in negotiating translation ambiguity. *Behavior Research Methods, 39*(4), 1029–1038.

Qualtrics. (2016). *Qualtrics research suite*. Provo, Utah, USA. http://www.qualtrics.com. Last retrieved May 6, 2017.

R Development Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria. http://www.r-project.org. Last retrieved May 6, 2017.

Rizopoulos, Dimitris. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1–25.

Simons, Gary F. & Fennig, Charles D. (2017). *Ethnologue: Languages of the world* (20th ed.). Dallas, TX: SIL International. http://www.ethnologue.com. Last retrieved May 6, 2017.

Talamas, Adrienne, Kroll, Judith F., & Dufour, Robert. (1999). From form to meaning: Stages in the acquisition of second-language vocabulary. *Bilingualism: Language and Cognition, 2*(1), 45–58.

Teng, Yanjiang. (2017). Hanyu Shuiping Kaoshi (HSK): Past, present, and future. In Zhang & Lin (Eds.), pp. 3–19.

Yap, Melvin J., Balota, David A., Tse, Chi-Shing, & Besner, Derek. (2008). On the additive effects of stimulus quality and word frequency in lexical decision: Evidence for opposing interactive influences revealed by RT distributional analyses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(3), 495–513.

Zhang, Dongbo & Lin, Chin-Hsi (Eds.). (2017). *Chinese as a second language assessment*. Singapore: Springer Nature Singapore Pte Ltd.

Proceedings of the 42nd annual Boston University Conference on Language Development

edited by Anne B. Bertolini
and Maxwell J. Kaplan

Cascadilla Press Somerville, MA 2018

ISSN 1080-692X
ISBN 978-1-57473-086-9 (2 volume set, paperback)
ISBN 978-1-57473-186-6 (2 volume set, library binding)

Ordering information

To order a copy of the proceedings or to place a standing order, contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, sales@cascadilla.com, www.cascadilla.com