

Shiwei Qi, Mark Feng Teng\* and Ailan Fu

# LexCH: a quick and reliable receptive vocabulary size test for Chinese Learners

<https://doi.org/10.1515/applirev-2022-0006>

Received January 21, 2022; accepted June 22, 2022; published online July 11, 2022

**Abstract:** The measurement of vocabulary size is crucial in applied linguistics research. Although increasing attention has been given to the study of Chinese vocabulary assessment, few reliable and valid tools are available to evaluate Chinese learners' receptive vocabulary size, particularly for teenagers and adults. We aim to fill this gap by developing LexCH, a quick, reliable and free receptive vocabulary size assessment tool that researchers and language teachers can readily adopt. In developing LexCH, we chose items covering a range of difficulty levels and with strong discriminative power as test items for the final version of LexCH based on item response theory. In total, 480 students from a junior high school and a high school in China participated in this study. Our initial validation results suggest that LexCH is a reliable and valid receptive vocabulary size test for L1 Chinese speakers; it also shows great potential for use among L2 Chinese learners. Implications for assessing receptive vocabulary size in Chinese learning are provided.

**Keywords:** Chinese vocabulary acquisition; item response theory (IRT); receptive vocabulary; vocabulary assessment

## 1 Introduction

Given the growing interest among scholars in applied linguistics and bilingual and multilingual studies, the development of quick, reliable, and preferably comparable assessment tools for understanding vocabulary size in different languages is essential. Vocabulary size is a sound predictor of many indices of linguistic ability, including reading (Anderson and Freebody 1983; Teng and Cui 2022), and listening (Stæhr 2009; Teng 2016). Many researchers have begun to use

---

**\*Corresponding author: Mark Feng Teng**, Center for Linguistic Sciences, Beijing Normal University, Zhuhai, China, E-mail: markteng@bnu.edu.cn. <https://orcid.org/0000-0002-5134-8504>

**Shiwei Qi and Ailan Fu**, Center for Linguistic Sciences, Beijing Normal University, Zhuhai, China, E-mail: 202031080011@mail.bnu.edu.cn (S. Qi), Ellenfu@bnu.edu.cn (A. Fu)

 Open Access. © 2022 the author(s), published by De Gruyter.  This work is licensed under the Creative Commons Attribution 4.0 International License.

language background surveys or to ask participants to rate their proficiency level on Likert scales to estimate language proficiency. A valid vocabulary size test could provide comparatively objective information about one's language proficiency—often quickly, conveniently, and free of charge (Amenta et al. 2021; Lemhöfer and Broersma 2012). In addition, evaluating vocabulary size is crucial in several research aspects: tracking vocabulary growth (e.g., Cheng et al. 2018; Keuleers et al. 2015); comparing different groups of speakers' vocabulary sizes (e.g., Farkas and Beron 2004); relating vocabulary size to language activities to inform classroom teaching, including viewing captioned videos (Teng 2021), developing teaching resources (Nation and Webb 2011), and matching learners with appropriate reading or listening materials (Nguyen and Nation 2011).

Although Chinese (Mandarin) is by far the world's most spoken first language (L1) (Eberhard et al. 2022), and an increasing number of people are studying Chinese as a second language (L2), a reliable and valid test to measure Chinese lexical size is lacking. A number of character recognition tests have been created (Chan and Chang 2018; Wen et al. 2015; Zhang et al. 2021), but they provide limited information about one's Chinese vocabulary size (as most Chinese words are composed of double and multiple characters). Meaning-recall interviews, in which participants are asked to orally define a group of chosen words, are time-consuming; their validity has also been called into question (Nation and Coxhead 2021). Given the need for and current absence of a reliable and valid vocabulary size test for Chinese learners, the present study seeks to provide Chinese language researchers with an easy-to-use test that makes it possible to measure Chinese vocabulary size in an objective and reliable way.

## 2 Literature review

### 2.1 Measuring vocabulary size

When assessing vocabulary knowledge, an important distinction concerns the *size* or *breadth* of vocabulary knowledge and the *depth* or *quality* of this knowledge (Anderson and Freebody 1981). Despite some debate over the conceptual difference between breadth and depth (e.g., Vermeer 2001), empirical studies have shown that the two are unique—at least for lower-frequency words and larger vocabulary sizes (Schmitt 2014). Most vocabulary size tests in this field assess the breadth of vocabulary knowledge, i.e., how many words are known, rather than the depth of vocabulary knowledge (Read 2000).

The Vocabulary Size Test (VST) (Nation and Beglar 2007), designed to measure L1 and L2 learners' written receptive vocabulary size in English, is one of the most

popular vocabulary size tests in applied linguistics. The sampling of test items is based on the frequency levels of word families occurring in the British National Corpus up to Level 6 (Bauer and Nation 1993). With the aim of capturing total vocabulary size, the VST “represents the various frequency levels without a bias towards any particular frequency levels” (Nation 2012, p.2). The test contains 140 items; they are presented in a decontextualized multiple-choice format, in which learners are instructed to choose the best definition of each word from four choices. Studies suggest that the VST has good validity (e.g., Beglar 2010). Even so, it has received criticism for its multiple-choice format because the construct of meaning recognition does not approximate real language activities such as reading and listening. Instead, the construct of meaning-recall, where learners must remember a word’s meaning upon seeing the term, is more closely related to everyday language use (Gyllstad et al. 2015; Kremmel and Schmitt 2016; Stewart 2014). A second alleged limitation of this format is that individuals’ scores can be inflated due to test strategies such as guessing (Nation and Coxhead 2021). Finally, the full VST typically takes between 40 and 60 min to complete, which is inefficient for research in which participants take multiple tests.

Another influential vocabulary size test is the Eurocentres Vocabulary Size Test (EVST) (Meara and Jones 1987, 1990). This test adopts a Yes/No format. The test was originally used to measure L1 vocabulary size (e.g., Anderson and Freebody 1983; Zimmerman et al. 1977) before later being adopted by second language researchers. In the EVST, items are presented as a list of words in isolation, and test takers are asked to respond “Yes” to the words they know and “No” to words they do not recognize. To avoid inflated scores, half of the test items are pseudowords. Points are deducted if test takers respond affirmatively to pseudowords. This kind of lexical judgement task is thought to represent a special type of meaning-recall task that is more similar to the mental processes accompanying reading and listening (Zhang et al. 2019). Moreover, test construction and administration are easier compared with other formats, making it possible to sample a high number of words and to collect data from large participant groups. Although the Yes/No format has been questioned in several aspects, further investigation has indicated that these issues are not truly problematic (Amenta et al. 2021). Some researchers have pointed out that in responding to a Yes/No item, learners can draw on partial knowledge of the word that is not strong enough to enable them to tackle the word when reading or listening (e.g., Nation and Coxhead 2021). However, others have reported that the results of Yes/No tests are closely correlated with other test types, such as multiple-choice, translation, and interview tasks (Nakata et al. 2020; Stubbe 2012; Zhang et al. 2019). Meanwhile, some scholars have critiqued the test’s isolated and decontextualized way of presenting items. However, we agree with Read (2000) that it is unnecessary to present words in context if an instrument is

not intended to measure test takers' ability to use contextual information when responding.

## 2.2 LexTALE

The Lexical Test for Advanced Learners of English (LexTALE) developed by Lemhöfer and Broersma (2012) is a vocabulary size test targeted at advanced English adult learners. It contains 60 items, including 40 words and 20 nonwords selected from the 240 items of an unpublished vocabulary size test called "10 K" by Meara (1996). Test takers are instructed to respond "Yes" to an existing English word and "No" to a nonexistent English word. LexTALE test results have been shown to correlate highly with general English proficiency (Lemhöfer and Broersma 2012; Nakata et al. 2020; Zhang et al. 2019). Given the sound validity and convenience of administering the LexTALE (which takes only 3–5 min to complete), researchers have adapted the instrument into Dutch and German (Lemhöfer and Broersma 2012), French (Lextale\_FR, Brysbaert 2013), Spanish (Lextale\_Esp, Izura et al. 2014), Chinese (LEXTALE\_CH, Chan and Chang 2018), and Italian (LexITA, Amenta et al. 2021). The Chinese version LEXTALE\_CH is a character-based vocabulary size test; that is, it evaluates how many characters test takers know. However, most Chinese words are combinations of multiple characters, so the number of known characters should not equal the number of words known. This consideration is detailed in the ensuing section (see Section 2.3).

All existing LexTALE versions adopt the same format and take word-frequency as the critical criterion. The actual number of items varies. LexTALE contains 40 words and 20 nonwords; Lextale\_FR consists of 56 words and 28 nonwords; Lextale\_Esp, LexITA, and LEXTALE\_CH have 60 words and 30 nonwords each. Raising the number of words is meant to increase the test's reliability and make the test applicable to both L1 and L2 speakers (Brysbaert 2013). Some researchers have contended that the test might not include enough difficult words for L1 speakers. For example, a study of LexITA reported a ceiling effect for the L1 group, with nearly perfect responses (mean score = 57.9/60) and a small standard deviation (SD = 2.15, range = 50–60 in Phase 1) (Amenta et al. 2021). The authors stated that the test was appropriate for L2 learners but should have increased the number of difficult words for L1 Italian speakers. Additionally, although Izura et al. (2014) argued in a study of Lextale\_Esp that there were no signs of a ceiling effect for L1 speakers, the average score for this group was 53.8 out of 60 (approximately 90%). Raising the number of difficult words could possibly lead to better discrimination among L1 speakers.

## 2.3 Measuring lexical knowledge in Chinese

Chinese characters are the basic unit of Chinese writing. Each character represents a syllable. In modern Chinese, some words are composed of only one character, such as 天 [sky], 火 [fire], and 跑 [run]. These words are called single-syllable simple words (Fu 2020) and constitute a small proportion of all Chinese words. Other Chinese words, including double- or multisyllable simple words, derivative words, and compound words, consist of more than one character. In many cases, the meaning of these words is not simply the combined meaning of each character. For instance, the double-syllable simple word 蝴蝶 [butterfly] is composed of 蝴 and 蝶, but neither character alone holds meaning (Fu 2020). Another example would be the compound word 大家 [everybody]. 大 means *big* and 家 means *family*, but 大家 means *everybody* rather than *big family* (Fu 2020). Therefore, the number of characters known in Chinese is not necessarily analogous to the number of words the person knows.

Yan et al. (2020) compared the predictive power of character-based vocabulary size and word-based vocabulary size in children's reading development in L1 Chinese. The importance of character-based vocabulary size in reading was found to decline with age, whereas the significance of word-based vocabulary size increased as children grew up. The development of a reliable word-based vocabulary size test is hence urgently needed. To date, researchers have mainly adopted two methods to measure L1 Chinese speakers' word-based receptive vocabulary size. The first is the oral definition task, where participants are required to orally define the given word (e.g., Chen et al. 2018; Chen et al. 2019; Cheng and Wu 2017; Li et al. 2012; Yan et al. 2020). However, native speakers sometimes struggle to explain the meanings of words despite knowing what these words mean. Test scoring can also be troublesome in this case (Nation and Coxhead 2021). Furthermore, the oral definition task usually occurs during one-on-one interviews. Collecting data from large samples is accordingly time-consuming. Another common assessment tool is the Chinese version of the Peabody Picture Vocabulary Test (PPVT) (Siu et al. 2015; Zhang 2017), in which test takers listen to several words and choose corresponding pictures. Despite the good reliability and validity of this test, it has been normed with children between ages 2.5 and 7.5 and may be too easy for older learners (Guo et al. 2019).

In the absence of a widely accepted receptive vocabulary size test for Chinese learners, the present study aims to develop and trial a word-based receptive vocabulary size test called LexCH. This paper attempts to answer the following research question:

Does the LexCH produce valid and reliable results?

### 3 Methodology

The LexCH was developed in line with steps used to create previous LexTale tests, particularly the Italian version LexITA (Amenta et al. 2021), the French version (Brysbaert 2013), and the Spanish version (Izura et al. 2014).

#### 3.1 Materials

Word items for LexCH were sampled from the *Lexicon of Common Words in Contemporary Chinese* (LCWCC, 现代汉语常用词表) released by the Department of Language Information Management of Ministry of Education of the PRC (Project Group of Lexicon Common Words in Contemporary Chinese 2008). LCWCC provides 56,008 “commonly used words in Putonghua” (p.1), referring to modern Mandarin vocabulary that Chinese L1 speakers with secondary education frequently encounter and use in daily life. The majority of these words are double-syllable words. Some common abbreviations, idioms, and other fixed phrases that express integral concepts are collected in the LCWCC,<sup>1</sup> but proper nouns and terms are not included.<sup>2</sup> All words are sorted by frequency.

We divided the words in LCWCC into 25 frequency levels, with each level containing approximately 2,200 words. Following Nation and Beglar (2007), the sampling of the word items for the initial version of LexCH represented the various frequency levels of the words without bias towards any particular level. In this way, the test items would better represent all words in LCWCC. Ten words were randomly chosen from each level. In total, 250 words were extracted from LCWCC, constituting test items for the initial version of LexCH. In addition, 75 nonwords were made by randomly selecting two characters from LCWCC to form a nonword in Chinese. We did not include nonwords that looked or sounded like real Chinese words to avoid confusing test takers (Amenta et al. 2021). We then searched for these words in the Chinese National Corpus and Google search engine to ensure the words’ nonexistence. We also asked five Ph.D. candidates majoring in applied linguistics to confirm that these nonwords did not make sense. The initial version of LexCH featured 250 words and 75 nonwords. These terms were then mixed, listed in a random order, and presented on paper. The final version of LexCH was presented in Appendix.

---

<sup>1</sup> Whether or not to include abbreviations, idioms, and phrases in vocabulary lists and tests is a controversial issue (Nation and Coxhead 2021). However, this consideration was not the focus of the present study, so we adhered to LCWCC’s word selection criteria.

<sup>2</sup> According to LCWCC, proper nouns and terms are not included in principle, unless the word has significantly high frequency and is regularly and socially used.

## 3.2 Participants

Two groups of participants took part in the present study. The first group consisted of 251 students (137 men, 107 women, and 7 who chose not to identify by sex) from a junior high school and a high school in Guangdong, China. Both were middle-ranking schools in the same city. The participants were from five intact classes between grades 7 and 11. This group of learners took the initial version of LexCH. The second group, who took the final version of LexCH, comprised 229 students from another 5 intact classes at the same schools (133 men, 82 women, and 14 who chose not to identify by sex). In total, 480 students participated in the study. Their mean age was 15 years ( $SD = 1.467$ ; range = 12–19; mode = 15). All participants spoke Chinese as their first language.

## 3.3 Procedure

The test was presented in a paper-and-pencil format. Signed consent forms were obtained from students' parents before the participants took the test. Additionally, students were asked to provide their personal information (e.g., age, grade, and sex) before taking the test. Their final grades in Chinese language classes were obtained from their school records. Next, students were instructed to review the list of words in LexCH and put a tick "✓" in the box if they knew the word. Participants were asked not to put a tick if they were unsure of the word's meaning. They were also told that the test results would not affect their course grades and that the test items included some nonwords (i.e., it would be pointless to put a tick for a word they did not know).

The first version of LexCH was piloted with a small group of learners similar to the target participants. The test instructions were changed to appear clear to the test takers. Then, the test papers were given to the teachers of the five classes who then distributed the test in class. To ensure that participants paid attention to the test instructions, the teachers read the directions aloud in front of the class before the test commenced. Furthermore, teachers ensured that all participants finished the test independently without referring to a dictionary. The test had no time limit, but all test takers ( $n = 251$ ) completed the initial version of LexCH (containing 325 items) within 15 min.

Based on data from the first version of LexCH, a two-parameter logistic model was run to determine the difficulty and discriminative power of each item. Items covering a range of difficulty and with strong discriminative power were chosen as test items in the final version of LexCH (144 items), which was then given to a new group of test takers ( $n = 229$ ).

### 3.4 Scoring

Following the Italian, Spanish, and French versions of the LexTALE test, the test scores were computed as follows:

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

This formula ensures that guessing behaviour is penalized. A test taker could obtain full points only if they answered “Yes” to all word items and “No” to all nonword items. If they responded “Yes” to all test items, they received zero points. Although there are more complicated ways to calculate scores, such as %Correct<sub>av</sub>,  $\Delta m$ , and  $I_{\text{SDT}}$  (Lemhöfer and Broersma 2012), Brysbaert’s (2013) study showed that these approaches produced similar results. The above formula represents the simplest calculation, and is convenient for researchers and teachers in classroom settings.

### 3.5 Data analysis

After collecting data for the first version of LexCH, we analysed the data using point-biserial correlation and item response theory (IRT). We opted for IRT because it accounts for test takers’ performance level and items’ level of difficulty. A two-parameter model of IRT can provide information about each item’s difficulty and discriminative power; as such, we were able to select items of varying difficulty and strong discriminative power to constitute the final version of LexCH. After that, following the validity framework proposed by Messick (1989, 1995), we provided initial validation of LexCH based on data from the second group of participants, who took the final version of the measure.

## 4 Results

### 4.1 Selecting items for LexCH

To evaluate the quality of each item in the initial LexCH, a point-biserial correlation between the response to each item and test takers’ overall accuracy was run separately on the 250 words and 75 nonwords. For a good-quality item, the point-biserial correlation coefficient should be positive, indicating that participants who claim to know this word are more likely to have higher overall accuracy. We first ran the correlation between the mean accuracy of each word item and the overall word item accuracy and then computed the correlation between the mean

accuracy of each nonword item and the total accuracy on nonword items. The correlation was positive for all words, which met our expectations.

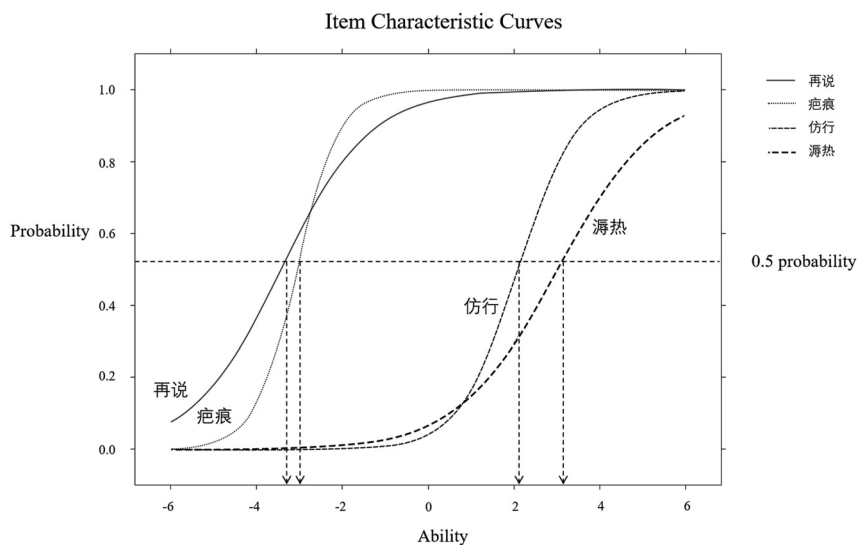
In addition, we performed IRT analysis on all test items because we wanted to select items that covered a range of difficulty levels and had strong discriminative power. IRT is a probabilistic model of how test takers respond to items based on the level of the underlying latent trait (i.e., vocabulary size in this study) (Desjardins and Bulut 2018). Within the IRT framework, items are characterized individually, and test characteristics are derived from the items. Characteristics in the two-parameter model include difficulty and discrimination; IRT assumes that these attributes are invariant across subgroups of examinees and across test administrations (Desjardins and Bulut 2018). Following Amenta and her colleagues (2021), we ran the two-parameter model on words and nonwords separately using the *mirt* R package (Chalmers 2012). First, we used the item fit function to examine the item fit statistics for item parameters from the two-parameter model. Ten words did not fit this model ( $p < 0.05$ ) and were removed from further analysis. Next, we computed the item characteristics for each item.

Figure 1 demonstrates an example of the item characteristic curves for four word items. In this graph, the difficulty value refers to the point on the x-axis where the item response curve crosses the 0.5 probability value on the y-axis. The item discrimination parameter is represented by the steepness of the response curve in its middle section: the steeper the curve is, the better the item's discriminative power. As seen in Figure 1, 再说 [besides] is easier than 疤痕 [scar] and 仿行 [replicate], and 潏热 [muggy] is most difficult. Additionally, 疤痕 has stronger discriminative power than 再说 and 仿行 is more discriminative than 潏热.

Based on each item's difficulty parameter, we ordered the word items and then chose them based on their discriminative power by extracting those with higher values at intervals of roughly one-twenty-fourth of the difficulty range. We therefore ensured that the items covered a range of difficulty levels and had strong discriminative power. Next, we repeated the same procedure for nonword items. As previous LexTALE tests were shown to be too easy for native or highly proficient speakers (Amenta et al. 2021), as suggested by Brysbaert (2013), we increased the number of items. The final version of LexCH included 96 words and 48 nonwords. We then assessed the test's reliability by computing the KR-20 coefficient, which reached 0.98 for all word items and 0.95 for all nonword items, reflecting strong reliability.

## 4.2 Validation

In this section, we aim to provide initial validity evidence for the final version of LexCH. We referred to Messick's (1989, 1995) framework of validity, which includes



**Figure 1:** Item response curves for four sample word items.

six aspects of construct validity that “... function as general validity criteria or standards for all educational and psychological measurement” (Messick 1995, p.6): content, substantive, structural, generalizability, external, and consequential validity. The first five aspects of construct validity are addressed below.

#### 4.2.1 Content aspect of construct validity

The content aspect of construct validity involves content relevance, representativeness, and technical quality (Messick 1989). Content relevance indicates the extent to which test items are related to the construct being measured, i.e., receptive knowledge of words’ form–meaning relationships. As mentioned earlier (see Section 2), we believe the Yes/No format of LexCH, which requires test takers to look at the written form of a word and recall its meaning, can effectively measure receptive vocabulary knowledge in terms of form–meaning connections. Moreover, LexCH is considered representative of the construct domain because (1) word items in the initial version of LexCH were selected based on a stratified random sampling method from 25 frequency levels, and (2) word items in the final version of LexCH were chosen according to their difficulty and discrimination value based on the two-parameter model (i.e., words with a wide range of difficulty levels and with strong discriminative power constituted test items in the final version of LexCH; Table 1). Technical quality can be investigated by inspecting

**Table 1:** Difficulty and discrimination estimate of word items on LexCH.

|                           | <i>N</i> | Range | Minimum | Maximum | Mean    | SD      |
|---------------------------|----------|-------|---------|---------|---------|---------|
| Difficulty                | 96       | 5.77  | −3.60   | 2.17    | −1.0122 | 1.23274 |
| Discrimination            | 96       | 2.53  | 1.27    | 3.80    | 2.1079  | 0.53169 |
| Valid <i>N</i> (listwise) | 96       |       |         |         |         |         |

item correlations and fit statistics. We believe the final version of LexCH has satisfactory technical quality for two reasons. First, point-biserial correlation analysis of the final version of LexCH showed that the correlation coefficients for nearly all items were positive (though two items had negative coefficients, accounting for approximately 1.4% of all test items); therefore, in general, more capable persons had higher overall accuracy. Second, fit analysis revealed that only four items showed misfit values ( $p < 0.05$ ), representing a mere 2.8% misfit rate. This outcome may be expected given the nature of the Type I error rate.

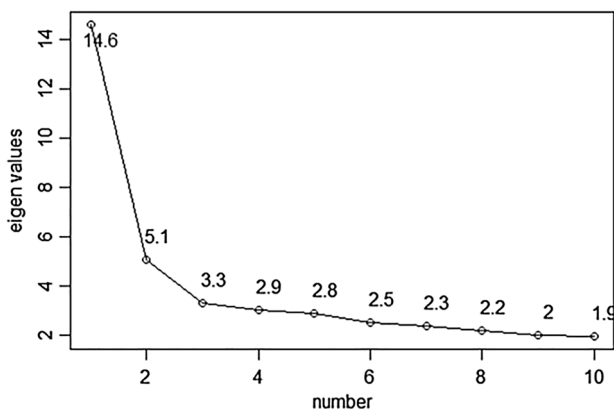
4.2.2 Substantive aspect of construct validity

The substantive aspect of construct validity can be determined with empirical evidence of response consistency or performance regularity reflective of domain processes (Loevinger 1957). This element can be evaluated by examining the consistency of each person’s response pattern with the item hierarchy (Webb et al. 2017). In the two-parameter model, *person fit* refers to the alignment between a test taker’s response pattern and the model. Person fit indices can be used to assess the validity of the selected model at the test taker level and the meaningfulness of the estimated latent trait levels (Embretson and Reise 2000). Therefore, we computed person fit statistics for all word items in the two-parameter model. A misfit person was defined as  $Z_h < -2$  (Desjardins and Bulut 2018). The misfit rate was 4.8% for the final version of LexCH, which is expected to occur by chance (less than 5%). Overall, test takers’ response patterns corresponded to the modelled difficulty order, providing supportive evidence for the substantive aspect of construct validity.

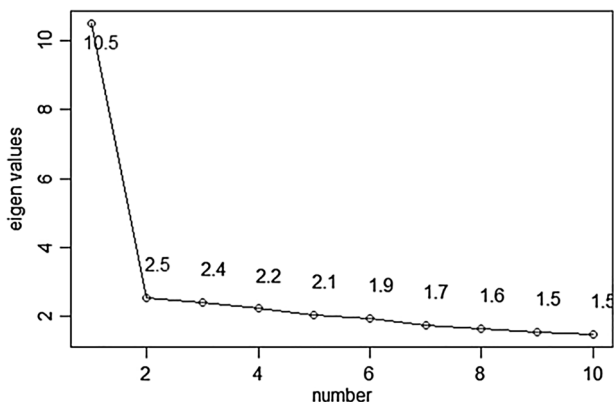
4.2.3 Structural aspect of construct validity

Regarding the structural aspect of construct validity, Messick (1989) argued that “... scoring models should be rationally consistent with what is known about the structural relations inherent in behavioural manifestations of the construct in question” and that “... the degree of homogeneity in the test should be commensurate with this characteristic degree of homogeneity associated with the construct”

(p.43). It was hypothesized that LexCH would show a high degree of psychometric unidimensionality for words and nonwords. Dimensionality can be tested via a scree plot (Desjardins and Bulut 2018). Plots were accordingly constructed for word and nonword items in this study. If only one major factor equal to the eigenvalues occurs immediately before the plot's elbow, then the unidimensionality is good (Desjardins and Bulut 2018). Figure 2 indicates one major factor and a second potential factor, as the eigenvalues began to level off thereafter. Because the eigenvalues of the first factor far outweighed those of the second factor, only one major factor appeared before the elbow. Word items displayed good unidimensionality as a result. Figure 3 depicts only one major factor before the elbow; the nonwords in LexCH therefore demonstrated good unidimensionality as well.



**Figure 2:** Scree plot of word item data (final version).



**Figure 3:** Scree plot of nonword item data (final version).

#### 4.2.4 Generalizability aspect of construct validity

Messick (1995) stated that the generalizability aspect examines “the extent to which score properties and interpretations generalise to and across population groups, settings, and tasks” (p.745). To discover the extent to which item and person calibrations were invariant across measurement contexts, a uniform differential item functioning (DIF) analysis was performed using the *dif* R package (Magis et al. 2010). More specifically, a DIF analysis was carried out with logistic regression to determine whether male and female participants demonstrated different probabilities of answering items correctly after being matched on measures of written receptive vocabulary knowledge. The results revealed that no statistically significant DIF applied to any word item ( $p < 0.05$ ). A single nonword item (Item 60) displayed significant DIF, representing only 0.7% of all items, which is expected to occur by chance (less than 5%).

The generalizability aspect of the construct aspect can also be considered based on the degree of reliability. We assessed the test’s reliability by computing the KR-20 coefficients for word items and nonword items. The KR-20 coefficient for word items in the final version of LexCH was 0.93 and that for nonword items was 0.92, suggesting that the final version was highly reliable. LexCH hence possessed generalizability.

#### 4.2.5 External aspects of construct validity

Messick (1989) pointed out that the external aspect refers to “the extent to which the test’s relationship with other tests and nontest behaviours reflects the expected high, low, and interactive relations implied in the theory of the construct being assessed” (p.45). Although examining participants’ performance on LexCH in relation to other tests measuring written receptive vocabulary knowledge would be useful, no widely accepted written receptive vocabulary size test is presently available for L1 Chinese speakers. Therefore, the external aspect of LexCH’s construct validity was examined in relation to participants’ age and Chinese literacy.

Age is a significant predictor of vocabulary size. In a massive study of native Dutch speakers’ vocabulary size, Keuleers and colleagues (2015) identified age as the most important factor affecting vocabulary size among the examined variables. Similarly, Coxhead et al. (2015) explored factors contributing to the vocabulary size of native speakers of English in New Zealand secondary schools. They found age to be the only factor that entered their regression model ( $p < 0.0005$ ,  $R^2 = 0.146$ ). As people age, they encounter more opportunities for language exposure, which creates productive conditions for vocabulary learning

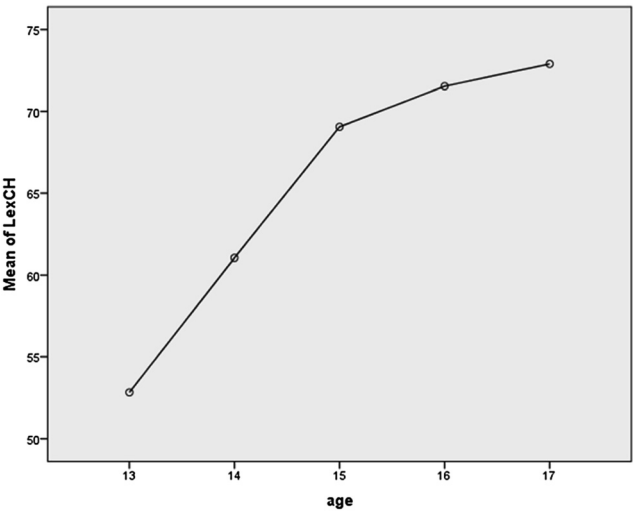


Figure 4: Mean scores on LexCH by age.

Table 2: Participants' LexCH scores by age.

| Age   | M.    | N   | SD     | Minimum | Maximum |
|-------|-------|-----|--------|---------|---------|
| 13    | 52.82 | 39  | 14.440 | 23      | 84      |
| 14    | 61.05 | 42  | 14.461 | 22      | 84      |
| 15    | 69.06 | 48  | 12.274 | 24      | 85      |
| 16    | 71.53 | 43  | 10.534 | 36      | 87      |
| 17    | 72.90 | 39  | 9.577  | 36      | 89      |
| Total | 65.68 | 211 | 14.331 | 22      | 89      |

(Nation and Coxhead 2021). Therefore, test takers' scores on LexCH were anticipated to rise with age.

As seen in Figure 4 and Table 2, the average LexCH scores (final version) rose with participants' age,<sup>3</sup> and the growth rate seemed to decelerate as age increased. This phenomenon is in line with Keuleers et al.'s (2015) study, which observed a slowdown in native vocabulary growth across age groups. We also performed a regression analysis to reveal the extent to which age could explain the variance in LexCH scores. The results highlighted that age was a significant

<sup>3</sup> As there were only three 12-year-olds, three 18-year-olds, and one 19-year-old (who could not represent their entire age group), their data were excluded from the age-related analysis in this section.

**Table 3:** Correlation coefficients between participants’ LexCH scores and Chinese literacy by grade.

| Grade    | <i>N</i> | Sig.  | <i>r</i>           |
|----------|----------|-------|--------------------|
| Seventh  | 40       | 0.006 | 0.431 <sup>b</sup> |
| Eighth   | 54       | 0.000 | 0.677 <sup>a</sup> |
| Ninth    | 34       | 0.012 | 0.424 <sup>c</sup> |
| Tenth    | 48       | 0.000 | 0.575 <sup>a</sup> |
| Eleventh | 52       | 0.005 | 0.387 <sup>b</sup> |

<sup>a</sup> $p < 0.001$ ; <sup>b</sup> $p < 0.01$ ; <sup>c</sup> $p < 0.05$ .

predictor of scores ( $p < 0.001$ ,  $F = 18.515$ ,  $R^2 = 0.264$ ), explaining 26.4% of the total variance, higher than the figure ( $R^2 = 0.146$ ) identified by Coxhead et al. (2015). This discrepancy is probably due to the difference in the unit of measurement: LexCH measured Chinese vocabulary size based on Chinese words, whereas Coxhead et al. (2015) assessed vocabulary size by the number of word families known. Thus, it makes sense that a gap would emerge between the total variances in vocabulary sizes in the two studies and that the explanatory power of age would differ.

Another aspect of interest in this study was the relationship between LexCH scores and students’ Chinese literacy. We expected receptive vocabulary knowledge to correlate significantly with test takers’ Chinese literacy, especially reading comprehension, as suggested in prior work (e.g., Li et al. 2012; Ouellette 2006; Peng et al. 2020; Yan et al. 2020). Given the current lack of an available standardized language proficiency test for L1 Chinese speakers, we obtained students’ final scores in the Chinese subject from school records as a measure of their Chinese literacy. A Pearson correlation analysis was conducted between LexCH scores and final scores in the Chinese subject by grade, as Chinese examination papers varied by grade level. Table 3 shows that participants in all grades had positive correlations between LexCH scores and Chinese literacy.

In sum, the relationships between LexCH scores and age and between LexCH scores and Chinese literacy reflected the expected relations in previous research. These patterns further substantiate the external aspect of construct validity.

## 5 Discussion

In this paper, we described the development and initial validation of LexCH, which closely followed the procedure adopted by Amenta et al. (2021), Brysbaert (2013), Izura et al. (2014) and Chan and Chang (2018). However, minor adjustments were

made to suit our target test takers. Table 4 presents a comparison between LexCH and other vocabulary assessment tools using a Yes/No format.

**Table 4:** Vocabulary assessment tools in Yes/No format.

| Assessment tool             | Target group             | Number of items                  | Item selection  | Instruction                                                                                                                                                                                                                                          |
|-----------------------------|--------------------------|----------------------------------|-----------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| EVST                        | L2 speakers              | Autoadaptive                     | Frequency-based | Learners are presented with a series of words, one that time, and are asked to indicate by pressing one of two keys on the computer whether they think they know that were not.                                                                      |
| LexTALE                     | Advanced L2 speakers     | 40 words and 20 nonwords         | Frequency-based | Indicate for each item whether it is an existing English word or not; respond no in case of doubt                                                                                                                                                    |
| Lextale_FR                  | L1 & L2 speakers         | 56 words and 28 nonwords         | IRT-based       | Indicate the words you know (or of which you are convinced are French words, even though you would not be able to give their precise meanings)                                                                                                       |
| Lextale_Esp                 | L1 & L2 speakers         | 60 words and 30 nonwords         | IRT-based       | Indicate the words you know (or of which you are convinced are Spanish words, even though you would not be able to give their precise meanings)                                                                                                      |
| LexITA                      | L2 speakers <sup>a</sup> | 60 words and 30 nonwords         | IRT-based       | Is this an Italian word?                                                                                                                                                                                                                             |
| LEXALE_CH (character-based) | L1 & L2 speakers         | 60 characters and 30 nonce items | IRT-based       | Check the box above the character if you identify it as an authentic Chinese character                                                                                                                                                               |
| LexCH (word-based)          | L1 speakers <sup>b</sup> | 96 words and 48 nonwords         | IRT-based       | Put a tick in the box if you know this word. Here “knowing a word” means you are able to understand the meaning of this word or are able to give a synonym for this word. If you are not sure about the meaning of a word, please do not put a tick. |

<sup>a</sup>Both L1 and L2 groups were included in the study, but LexITA was reportedly most suitable for L2 speakers (Amenta et al. 2021). <sup>b</sup>LexCH may potentially be used with L1 and L2 speakers, but has been only validated with L1 speakers thus far.

First, in terms of the number of test items, LexCH contained 96 words and 48 nonwords more than in previous LexTALE versions. As suggested in the study of LexITA (Amenta et al. 2021), the test showed a ceiling effect for L1 speakers. We therefore aimed to increase the number of items to better discriminate among Chinese learners. The findings showed that the mean accuracy for our participants was approximately 68%, lower than that for LexITA (97%), Lextale\_Esp (90%) and Lextale\_FR (88%) for L1 groups. The standard deviation was 14.331, which clearly avoided any ceiling effect. Second, regarding the unit of measurement, LexCH is a word-based vocabulary test unlike LEXTALE\_CH. As mentioned, the number of characters known does not equal the number of words known. For Chinese learners with secondary education, a character-based vocabulary test may not be sufficiently discriminative, as there are only 3,500 common characters in Chinese. We maintain that LexCH would be more suitable for Chinese learners, particularly teenagers and adults. Third, other LexTALE tests asked test takers to decide whether items were English/Italian/Spanish/French words or authentic Chinese characters. Test takers were instructed to respond “Yes” if they thought the items were real words or characters, even if they did not know items’ precise meanings (except for the English version LexTALE, in which learners were instructed to answer “No” in case of doubt). In LexCH, we did not ask participants “Is this a Chinese word?”, and this was determined by the characteristics of Chinese words. The Chinese language contains a large number of compound words. Deciding whether two or more characters form a word is a controversial linguistic concern (Fu 2020). Furthermore, the Yes/No format of assessing receptive vocabulary knowledge has been criticized for asking participants to recognize a word form, because form-recognition alone is not sufficient for linguistic comprehension (Stoeckel et al. 2020). Therefore, we asked test takers whether they knew the meaning of provided words. They were not encouraged to respond affirmatively if they were unsure about a word’s meaning. In this way, the strength of knowledge that LexCH measures is supposed to be stronger than that of other LexTALE tests, and the test format is more similar to meaning-recall rather than to form-recognition. The test results should hence arguably be a better predictor of language proficiency, particularly reading comprehension (Nation and Coxhead 2021).

One limitation of the existing receptive vocabulary size tests lies in the representativeness of a small, random sample of target words (Stoeckel et al. 2020). Many vocabulary size assessments sampled test items based on frequency bands (e.g., Meara and Jones 1990; Nation and Beglar 2007; Webb et al. 2017; Schmitt et al. 2001), assuming that the items within a frequency-derived word band

featured a similar level of difficulty. However, empirical evidence indicates that difficulty can vary significantly for individual words within frequency bands (Beglar 2010) and that frequency alone explains only 25% of the variance in word recognition difficulty (Hashimoto and Egbert 2019). To address this issue, when constructing LexCH, we adopted a stratified random sampling method to obtain 250 words from 25 frequency levels to form the initial version. We then employed the IRT-based method to choose words with a range of difficulty and strong discriminative power to develop the final version of the assessment. This procedure significantly improved the test items' representativeness. Word items were sampled from the *LCWCC*, which contained 56,008 “commonly used words in Putonghua” (p.1). LexCH is thus thought to be most suitable for L1 Chinese speakers with secondary education and perhaps for Chinese L2 learners, especially advanced groups. We do not intend for LexCH to be used among native speakers of Chinese with a higher education background because *LCWCC* does not contain academic terms that university students or graduates of different disciplines are likely to know, even though these learners could show some variation in LexCH performance. Additionally, we do not recommend LexCH for use with small children because the Yes/No format requires a high degree of judgement. Use with this age group could significantly threaten the validity of the instrument.

## 6 Limitations and implications

Our initial validation provides supportive evidence for five aspects of construct validity proposed by Messick (1995). Vocabulary size measured by LexCH correlated significantly with Chinese literacy ( $0.387 \leq r \leq 0.677$ ,  $p < 0.05$ ). However, one should be cautious when interpreting this result, as one limitation of our study is that we took students' final scores in the Chinese subject at school as measures of their Chinese literacy. Their exams differed by grade and had not been validated. Further empirical studies are needed before the vocabulary size measured by LexCH can certifiably be used as a rough measure of Chinese proficiency for L1 Chinese speakers in psychological and linguistic research or placement tests.

A second limitation of the current study entails the lack of evidence of criterion validity. To the best of our knowledge, no available and widely recognized receptive vocabulary size test yet exists for teenage and adult learners of

Chinese. Researchers could collect evidence in the future through meaning-recall interviews and other means to further validate this test.

Third, for practical reasons, test items were presented in a fixed order on paper. We do not consider these parameters problematic. Izura et al.' (2014) suggested that a fixed presentation order can still produce good results. However, test items could be administered through a random permutation via computer in subsequent studies to avoid potential effects from items' positions.

Finally, we provided validity evidence based only on data collected from L1 speakers. Our sample was also fairly homogeneous in that all participants were teenage students from one junior high school and one high school in a city in Guangdong, China. Further validation efforts are needed with L1 speakers of different ages and from diverse educational backgrounds as well as with L2 speakers at various proficiency levels.

## 7 Conclusion

Overall, the present study frames LexCH as a quick and reliable assessment tool for Chinese receptive vocabulary size among L1 Chinese speakers. The test also shows promise for use with L2 learners. The process through which we developed LexCH mirrored methods employed with previous LexTALE tests. The procedures were slightly adapted to fit the characteristics of Chinese words and the needs of our target test takers. Researchers interested in bilingual or multilingual processes can utilize this measure to simply and consistently assess individuals' Chinese vocabulary size. LexCH takes approximately 7 min to complete, making it an ideal complement to other tests as required by some research designs. Despite the encouraging evidence of validity at this stage, LexCH must be explored further with groups of learners other than those described in this study. Multiple validation approaches should also be deployed to investigate other aspects of validity in the future.

**Acknowledgments:** We would like to extend special thanks to Dr. Liu Hao and Mr. Chen Xi from Beijing Normal University for their advice on the data analysis in this study. We are also grateful to anonymous reviewers for their kind and useful suggestions for improving this article.

**Conflict of interest statement:** The authors declare that there is no conflict of interest involved in this article.

**Availability of data and material:** The data and materials that support the findings of this study are available on request from the corresponding author.

Appendix

中文词汇量测试

姓名：\_\_\_\_\_ 学校：\_\_\_\_\_ 班级：\_\_\_\_\_ 年龄：\_\_\_\_\_ 性别：\_\_\_\_\_

测试说明：本测试的目的为测试您的中文词汇量，所收集数据严格保密，仅作科研用途，不会影响您的在校成绩，请放心如实作答！

请在以下词语中给您认识的词打“√”，此处“认识”是指您能够理解该词的意思或给出该词的近义词。如果您不确定该词的意思，则请勿勾选。

请注意：以下词汇中包含一定量的假词，即实际上不存在的词。如果您勾选了假词，则需要倒扣分。因此，请您根据实际情况认真作答！

| 词         | 您是否认识该词？ | 词        | 您是否认识该词？ |
|-----------|----------|----------|----------|
| 1. 爱国     |          | 32. 解气   |          |
| 2. 现有     |          | 33. 端架子  |          |
| 3. 祖居     |          | 34. 悬圈   |          |
| 4. 坦油     |          | 35. 作巴   |          |
| 5. 嗷嗷待哺   |          | 36. 预演   |          |
| 6. 拨臊     |          | 37. 兼管   |          |
| 7. 秋毫     |          | 38. 坝久   |          |
| 8. 深幽     |          | 39. 社会福利 |          |
| 9. 悍然     |          | 40. 管教所  |          |
| 10. 期望    |          | 41. 明宜   |          |
| 11. 好歹    |          | 42. 公厕   |          |
| 12. 逢     |          | 43. 接触   |          |
| 13. 油索    |          | 44. 打躬作揖 |          |
| 14. 当天    |          | 45. 巧计   |          |
| 15. 惨兴    |          | 46. 赤庆   |          |
| 16. 电子计算机 |          | 47. 小家伙  |          |
| 17. 相依    |          | 48. 强梁   |          |
| 18. 照料    |          | 49. 摔释   |          |
| 19. 深明大义  |          | 50. 菜卫   |          |
| 20. 棋炎    |          | 51. 驯兽   |          |
| 21. 改写    |          | 52. 引以为荣 |          |
| 22. 洽挂    |          | 53. 时期   |          |
| 23. 兑取    |          | 54. 吻合   |          |
| 24. 疏伟    |          | 55. 叩乔   |          |
| 25. 跃动    |          | 56. 称贺   |          |
| 26. 责问    |          | 57. 南作   |          |
| 27. 奶油小生  |          | 58. 义父   |          |
| 28. 斑营    |          | 59. 技法   |          |
| 29. 发目    |          | 60. 博涤   |          |
| 30. 增收节支  |          | 61. 通复   |          |
| 31. 不谋而合  |          | 62. 良现   |          |

(continued)

| 词        | 您是否认识该词？ | 词         | 您是否认识该词？ |
|----------|----------|-----------|----------|
| 63. 挤压   |          | 105. 唇成   |          |
| 64. 玉超   |          | 106. 乙雅   |          |
| 65. 君子   |          | 107. 决策   |          |
| 66. 时笔   |          | 108. 掌鞭   |          |
| 67. 株连   |          | 109. 疤痕   |          |
| 68. 魔窟   |          | 110. 仿行   |          |
| 69. 子爵   |          | 111. 胶劳   |          |
| 70. 单纯   |          | 112. 西化   |          |
| 71. 推进   |          | 113. 闷指   |          |
| 72. 赘肉   |          | 114. 赘牙   |          |
| 73. 锅炉   |          | 115. 歪程   |          |
| 74. 押车   |          | 116. 枪刺   |          |
| 75. 评剧   |          | 117. 导断   |          |
| 76. 锐办   |          | 118. 庄园主  |          |
| 77. 豪经   |          | 119. 意象   |          |
| 78. 中高档  |          | 120. 百叶   |          |
| 79. 典范   |          | 121. 永生永世 |          |
| 80. 周椅   |          | 122. 摇摆   |          |
| 81. 锣鼓喧天 |          | 123. 桔反   |          |
| 82. 串黛   |          | 124. 制作   |          |
| 83. 挥泪   |          | 125. 儿女情长 |          |
| 84. 心血管  |          | 126. 人数   |          |
| 85. 主口   |          | 127. 极育   |          |
| 86. 工辩   |          | 128. 福寿   |          |
| 87. 野野   |          | 129. 头纱   |          |
| 88. 高知   |          | 130. 排列   |          |
| 89. 纳税   |          | 131. 并早   |          |
| 90. 优东   |          | 132. 检测   |          |
| 91. 重量   |          | 133. 成对   |          |
| 92. 炸裂   |          | 134. 笑嘻嘻  |          |
| 93. 巨录   |          | 135. 火药味  |          |
| 94. 仗脆   |          | 136. 显示器  |          |
| 95. 当权派  |          | 137. 小仓库  |          |
| 96. 悬绒   |          | 138. 感悟   |          |
| 97. 底样   |          | 139. 吹嘘   |          |
| 98. 时而   |          | 140. 呱呱坠地 |          |
| 99. 互敬互爱 |          | 141. 缩杰   |          |
| 100. 号称  |          | 142. 聚钮   |          |
| 101. 正告  |          | 143. 拦河坝  |          |
| 102. 子口  |          | 144. 退根   |          |
| 103. 奋光  |          |           |          |
| 104. 蒙怀  |          |           |          |

## English Translated Version

### LexCH: Lexical Test for Chinese Learners

Name: \_\_\_\_\_ School: \_\_\_\_\_ Class: \_\_\_\_\_ Age: \_\_\_\_\_ Sex: \_\_\_\_\_

Instruction: The purpose of the test is to measure your Chinese vocabulary size. The data collected will be confidential and only used for research purposes. The results of this test will not affect your academic performance at school, so please respond to the questions honestly.

**Please put a tick “√” in the box if you know this word. Here “knowing a word” means you are able to understand the meaning of this word or are able to give a synonym for this word. If you are not sure about the meaning of a word, please do not put a tick.**

**Attention: Some of the following words are nonwords, which means they do not really exist. If you tick the nonwords, points will be deducted. So please take the test seriously.**

| word      | Do you know this word? | word     | Do you know this word? |
|-----------|------------------------|----------|------------------------|
| 1. 爱国     |                        | 32. 解气   |                        |
| 2. 现有     |                        | 33. 端架子  |                        |
| 3. 祖居     |                        | 34. 悬圈   |                        |
| 4. 坦油     |                        | 35. 作巴   |                        |
| 5. 嗷嗷待哺   |                        | 36. 预演   |                        |
| 6. 拨躁     |                        | 37. 兼管   |                        |
| 7. 秋毫     |                        | 38. 坝久   |                        |
| 8. 深幽     |                        | 39. 社会福利 |                        |
| 9. 悍然     |                        | 40. 管教所  |                        |
| 10. 期望    |                        | 41. 明宣   |                        |
| 11. 好歹    |                        | 42. 公厕   |                        |
| 12. 逢     |                        | 43. 接触   |                        |
| 13. 油索    |                        | 44. 打躬作揖 |                        |
| 14. 当天    |                        | 45. 巧计   |                        |
| 15. 惨兴    |                        | 46. 赤庆   |                        |
| 16. 电子计算机 |                        | 47. 小家伙  |                        |
| 17. 相依    |                        | 48. 强梁   |                        |
| 18. 照料    |                        | 49. 摔释   |                        |
| 19. 深明大义  |                        | 50. 菜卫   |                        |
| 20. 棋炎    |                        | 51. 驯兽   |                        |
| 21. 改写    |                        | 52. 引以为荣 |                        |
| 22. 洽挂    |                        | 53. 时期   |                        |
| 23. 兑取    |                        | 54. 吻合   |                        |
| 24. 疏伟    |                        | 55. 叩乔   |                        |
| 25. 跃动    |                        | 56. 称贺   |                        |
| 26. 责问    |                        | 57. 南作   |                        |
| 27. 奶油小生  |                        | 58. 义父   |                        |
| 28. 斑营    |                        | 59. 技法   |                        |
| 29. 发目    |                        | 60. 博涤   |                        |
| 30. 增收节支  |                        | 61. 通复   |                        |
| 31. 不谋而合  |                        | 62. 良现   |                        |

(continued)

| word     | Do you know this word? | word      | Do you know this word? |
|----------|------------------------|-----------|------------------------|
| 63. 挤压   |                        | 105. 脍炙   |                        |
| 64. 玉超   |                        | 106. 乙雅   |                        |
| 65. 君子   |                        | 107. 决策   |                        |
| 66. 时笔   |                        | 108. 掌鞭   |                        |
| 67. 株连   |                        | 109. 疤痕   |                        |
| 68. 魔窟   |                        | 110. 仿行   |                        |
| 69. 子爵   |                        | 111. 胶劳   |                        |
| 70. 单纯   |                        | 112. 西化   |                        |
| 71. 推进   |                        | 113. 闷指   |                        |
| 72. 赘肉   |                        | 114. 象牙   |                        |
| 73. 锅炉   |                        | 115. 歪程   |                        |
| 74. 押车   |                        | 116. 枪刺   |                        |
| 75. 评剧   |                        | 117. 导断   |                        |
| 76. 锐办   |                        | 118. 庄园主  |                        |
| 77. 豪经   |                        | 119. 意象   |                        |
| 78. 中高档  |                        | 120. 百叶   |                        |
| 79. 典范   |                        | 121. 永生永世 |                        |
| 80. 周椅   |                        | 122. 摇摆   |                        |
| 81. 锣鼓喧天 |                        | 123. 桔反   |                        |
| 82. 串黛   |                        | 124. 制作   |                        |
| 83. 挥泪   |                        | 125. 儿女情长 |                        |
| 84. 心血管  |                        | 126. 人数   |                        |
| 85. 主口   |                        | 127. 极育   |                        |
| 86. 工辩   |                        | 128. 福寿   |                        |
| 87. 鼯野   |                        | 129. 头纱   |                        |
| 88. 高知   |                        | 130. 排列   |                        |
| 89. 纳税   |                        | 131. 并早   |                        |
| 90. 优东   |                        | 132. 检测   |                        |
| 91. 重量   |                        | 133. 成对   |                        |
| 92. 炸裂   |                        | 134. 笑嘻嘻  |                        |
| 93. 巨录   |                        | 135. 火药味  |                        |
| 94. 仗脆   |                        | 136. 显示器  |                        |
| 95. 当权派  |                        | 137. 小仓库  |                        |
| 96. 悬绒   |                        | 138. 感悟   |                        |
| 97. 底样   |                        | 139. 吹嘘   |                        |
| 98. 时而   |                        | 140. 呱呱坠地 |                        |
| 99. 互敬互爱 |                        | 141. 缩杰   |                        |
| 100. 号称  |                        | 142. 聚钮   |                        |
| 101. 正告  |                        | 143. 拦河坝  |                        |
| 102. 子口  |                        | 144. 退根   |                        |
| 103. 奋光  |                        |           |                        |
| 104. 蒙怀  |                        |           |                        |

## How to score LexCH

To score LexCH simply follow this formula

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

Use the following list to identify word and nonword items:

Word items:

1, 2, 3, 5, 7, 8, 9, 10, 11, 12, 14, 16, 17, 18, 19, 21, 23, 25, 26, 27,  
30, 31, 32, 33, 36, 37, 39, 40, 42, 43, 44, 45, 47, 48, 51, 52, 53, 54,  
56, 58, 59, 63, 65, 67, 68, 69, 70, 71, 72, 73, 74, 75, 78, 79, 81, 83,  
84, 88, 89, 91, 92, 95, 97, 98, 99, 100, 101, 102, 107, 108, 109, 110,  
112, 114, 116, 118, 119, 120, 121, 122, 124, 125, 126, 128, 129, 130,  
132, 133, 134, 135, 136, 137, 138, 139, 140, 143

Nonword items:

4, 6, 13, 15, 20, 22, 24, 28, 29, 34, 35, 38, 41, 46, 49, 50, 55, 57  
60, 61, 62, 64, 66, 76, 77, 80, 82, 85, 86, 87, 90, 93, 94, 96, 103  
104, 105, 106, 111, 113, 115, 117, 123, 127, 131, 141, 142, 144

## References

- Amenta, Simona, Linda Badan & Marc Brysbaert. 2021. LexITA: A quick and reliable assessment tool for Italian L2 receptive vocabulary size. *Applied Linguistics* 42(2). 292–314.
- Anderson, Richard & Peter Freebody. 1981. Vocabulary knowledge. In J. T. Guthrie (ed.), *Comprehension and teaching: Research reviews*, 77–117. Newark, DE: International Reading Association.
- Anderson, Richard & Peter Freebody. 1983. Reading comprehension and the assessment and acquisition of word knowledge. In B. Huxton (ed.), *Advances in reading/language research*, 2, 231–256. Greenwich, Connecticut: JAI Press.
- Bauer, Laufer & Paul Nation. 1993. Word families. *International Journal of Lexicography* 6(4). 253–279.
- Beglar, David. 2010. A Rasch-based validation of the vocabulary size test. *Language Testing* 27(1). 101–118.
- Brysbaert, Marc. 2013. A fast, free, and efficient test to measure language proficiency in French. *Psychologica Belgica* 53(1). 23–37.
- Chalmers, R. Philip. 2012. mirt: A multidimensional Item Response Theory package for the R environment. *Journal of Statistical Software* 48(6). <https://doi.org/10.18637/jss.v048.i06>.
- Chan, I. Lei & Charles. Chang. 2018. LEXTALE\_CH: A quick, character-based proficiency test for Mandarin Chinese. *Proceedings of the 42nd Annual Boston University Conference on language development*, 114–130. Somerville, MA: Cascadilla Press.

- Chen, Hongjun, Ying Zhao, Xinchuan Wu, Peng Sun, Ruibo Xie & Jie Feng. 2019. The relation between vocabulary knowledge and reading comprehension in Chinese elementary children: A cross-lagged study. *Acta Psychology Sinica* 51(8). 924–934.
- Chen, Jing, Tzu-Jung Lin, Yu-Min Ku, Jie Zhang & Ann O'Connell. 2018. Reader, word, and character attributes contributing to Chinese children's concept of word. *Scientific Studies of Reading* 22(3). 209–224.
- Cheng, Yahua & Xinchun Wu. 2017. The relationship between SES and reading comprehension in Chinese: A mediation model. *Frontiers in Psychology* 8. <https://doi.org/10.3389/fpsyg.2017.00672>.
- Cheng, Yahua, Xinchun Wu, Hongyun Liu & Hong Li. 2018. The developmental trajectories of oral vocabulary knowledge and its influential factors in Chinese primary school students. *Acta Psychology Sinica* 50(2). 206–215.
- Coxhead, Averil, Nation Paul & Dalice Sim. 2015. The vocabulary size of native speakers of English in New Zealand secondary schools. *New Zealand Journal of Educational Studies* 50(1). 121–135.
- Desjardins, Christopher D. & Bulut Okan. 2018. *Handbook of educational measurement and psychometrics using R*. Boca Raton: Chapman and Hall/CRC Press.
- Eberhard, David, Gary Simons & Charles Fennig (eds.). 2022. *Ethnologue: Languages of the world*, 25th edn. Dallas, Texas: SIL International. Online version. <http://www.ethnologue.com>.
- Embretson, Susan E. & Steven P. Reise. 2000. *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Farkas, George & Kurt Beron. 2004. The detailed age trajectory of oral vocabulary knowledge: Differences by class and race. *Social Science Research* 33(3). 464–497.
- Fu, Huaqing. 2020. *Xian Dai Han Yu Ci Hui [Modern Chinese Lexis]*. Beijing: Peking University Press.
- Guo, Xuan, Yuanlai Zhu, Xiaoyan Jiao, Bailing Zhang & Molidahan Aierken. 2019. Validity and reliability of the Chinese and Kazakh versions of the Peabody picture vocabulary test-forth edition in preschool children. *Chinese Mental Health Journal* 33(11). 845–850.
- Gyllstad, Henrik, Laura Vilkaite & Norbert Schmitt. 2015. Assessing vocabulary size through multiple-choice formats. *International Journal of Applied Linguistics* 166(2). 278–306.
- Hashimoto, Brett J. & Jesse Egbert. 2019. More than frequency? Exploring predictors of word difficulty for second language learners. *Language Learning* 69(4). 839–872.
- Izura, C., F. Cuetos & M. Brysbaert. 2014. Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicología* 35(1). 49–66.
- Keuleers, Emmanuel, Michaël Stevens, Paweł Mandera & Brysbaert Marc. 2015. Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology* 68(8). 1665–1692.
- Kremmel, Benjamin & Norbert Schmitt. 2016. Interpreting vocabulary test scores: What so various item formats tell us about learners' ability to employ words? *Language Assessment Quarterly* 13(4). 377–392.
- Lemhöfer, Kristin & Mirjam Broersma. 2012. Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods* 44(2). 325–343.
- Li, Tong, Catherine McBride-Chang, Anita Wong & Hua Shu. 2012. Longitudinal predictors of spelling and reading comprehension in Chinese as an L1 and English as an L2 in Hong Kong Chinese children. *Journal of Educational Psychology* 104(2). 286–301.
- Loevinger, Jane. 1957. Objective tests as instruments of psychological theory [Monograph]. *Psychological Reports* 3. 635–694.

- Magis, David, Sébastien Beland, Francis Tuerlinckx & Paul De Boeck. 2010. A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods* 42. 547–862.
- Meara, Paul. 1996. *English vocabulary tests: 10 k. Unpublished manuscript*. Swansea: Center for Applied Language Studies.
- Meara, Paul & Gary Jones. 1987. Test of vocabulary size in English as a foreign language. *Polyglotte* 8. 1–40.
- Meara, Paul & Gary Jones. 1990. *Eurocentres vocabulary size test 10 KA*. Zurich: Eurocentres.
- Messick, Samuel. 1989. Educational measurement. In R. Linn (ed.), *Validity*, 13–103. New York: NY: Macmillan.
- Messick, Samuel. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist* 50(9). 741–749.
- Nakata, Tatauya, Yu Tamura & Scott Aubrey. 2020. Examining the validity of the LexTALE test for Japanese college students. *The Journal of Asia TEFL* 17(2). 335–348.
- Nation, Paul. 2012. The vocabulary size test. <https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests/the-vocabulary-size-test/Vocabulary-Size-Test-information-and-specifications.pdf> (accessed 20 April 2022).
- Nation, Paul & David Beglar. 2007. A vocabulary size test. *The Language Teacher* 31(7). 9–13.
- Nation, Paul & Averil Coxhead. 2021. *Measuring native-speaker vocabulary size*. Amsterdam, Netherlands: John Benjamins B.V.
- Nation, Paul & Stuart Webb. 2011. *Researching and analysing vocabulary*. Boston, MA: Heinle, Cengage Learning.
- Nguyen, Le Thi Cam & Paul Nation. 2011. A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal* 42(1). 86–99.
- Ouellette, Gene. 2006. What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology* 98(3). 554–566.
- Peng, Peng, Kejin Lee, Jie Luo, Shuting Li, Malatesha Joshi & Sha Tao. 2020. Simple view of reading in Chinese: A one-stage meta-analytic structural equation modeling. *Review of Educational Research* 91(1). 1–33.
- Project Group of Lexicon Common Words in Contemporary Chinese. 2008. *Lexicon of common words in Contemporary Chinese*. Beijing: The Commercial Press.
- Read, John. 2000. *Assessing vocabulary*. Cambridge: Cambridge University Press.
- Schmitt, Norbert, Diane Schmitt & Caroline Clapham. 2001. Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing* 18(1). 55–88.
- Schmitt, Norbert. 2014. Size and depth of vocabulary knowledge: What the research shows. *Language Learning* 64(4). 913–951.
- Siu, Tik-Sze & Suk-Han Ho. 2015. Cross-language transfer of syntactic skills and reading comprehension among young Cantonese-English bilingual students. *Reading Research Quarterly* 50(3). 313–336.
- Stæhr, Lars Stenius. 2009. Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition* 31(4). 577–607.
- Stewart, Jeffrey. 2014. Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly* 11(3). 271–282.
- Stoeckel, Tim, Stuart McLean & Paul Nation. 2020. Limitations of size and levels tests or written receptive vocabulary knowledge. *Studies in Second Language Acquisition* 43(1). 181–203.

- Stubbe, Raymond. 2012. Do pseudoword false alarm rates and overestimation rates in Yes/No vocabulary tests change with Japanese university students' English ability levels? *Language Testing* 29(4). 471–488.
- Teng, Feng. 2016. An in-depth investigation into the relationship between vocabulary knowledge and academic listening comprehension. *TESL-EJ* 20(2) 1–17.
- Teng, Feng. 2021. *Language learning through captioned videos: Incidental vocabulary acquisition*. New York: Routledge.
- Teng, Feng & Yachong Cui. 2022. *The role of vocabulary knowledge, morphological awareness, and working memory in reading comprehension* [Manuscript submitted for publication]. Zhuhai: Beijing Normal University.
- Vermeer, Anne. 2001. Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics* 22(2). 217–234.
- Webb, Stuart, Yosuke Sasao & Oliver Balance. 2017. The updated vocabulary levels test. *ITL – International Journal of Applied Linguistics* 168(1). 33–69.
- Wen, Hong, Wenjun Tang & Xianwei Liu. 2015. Design of a test on quantity of literacy for students in the stage of compulsory education. *Yu Yan Wen Zi Ying Yong* 3. 88–100.
- Yan, Mengge, Hong Li, Yixun Li, Xielian Zhou, Yi Hui, Yahua Cheng & Xinchun Wu. 2020. The importance of decoding skills and vocabulary to reading comprehension in Chinese reading development. *Psychological Development and Education* 36(3). 311–317.
- Zhang, Dongbo. 2017. Multidimensionality of morphological awareness and text comprehension among young Chinese readers in a multilingual context. *Learning and Individual Differences* 56. 13–23.
- Zhang, Haiwei., Xueyan Zhang, Tiejun Zhang & Ruixin Wang. 2021. The creation and validation of a Hanzi recognition size test for learners of Chinese as a second language. *Chinese Teaching in the World* 35(1). 126–142.
- Zhang, Xian, Jianda Liu & Haiyang Ai. 2019. Pseudowords and guessing in the Yes/No format vocabulary test. *Language Testing* 37(1). 6–30.
- Zimmerman, Joel, Paul Broder, John Shaughnessy & Benton Underwood. 1977. A recognition test of vocabulary using signal-detection measures, and some correlates of word and nonword recognition. *Intelligence* 1(1). 5–31.

## Bionotes

### Shiwei Qi

Center for Linguistic Sciences, Beijing Normal University, Zhuhai, China  
202031080011@mail.bnu.edu.cn

Shiwei Qi obtained her MA in Teaching English to Speakers of Other Languages (TESOL) from King's College London and is currently a Ph.D. candidate in applied linguistics at Center for Linguistic Sciences, Beijing Normal University. Her research interests include vocabulary learning and language assessment.

**Mark Feng Teng**

Center for Linguistic Sciences, Beijing Normal University, Zhuhai, China

**markteng@bnu.edu.cn**

**<https://orcid.org/0000-0002-5134-8504>**

Mark Feng Teng obtained his Ph.D. from Hong Kong Baptist University and is currently Associate Professor of Applied Linguistics at Beijing Normal University. His research interests mainly focus on L2 vocabulary acquisition, and metacognition in L2 writing. He has published extensively in those areas, including monographs and edited books, numerous international journal papers, and book chapters. He received the 2017 Hong Kong Association for Applied Linguistics (HAAL) Best Paper Award. He serves as co-editor for OTI section, TESL-EJ.

**Ailan Fu**

Center for Linguistic Sciences, Beijing Normal University, Zhuhai, China

**Ellenfu@bnu.edu.cn**

Ailan Fu is a Professor of Linguistics at Center for Linguistic Sciences, Beijing Normal University. Her research interests include Sino-Tibetan linguistics, teaching Chinese as a second language, language testing and assessment, and Mandarin teaching in Hong Kong. She has published 6 academic books and 8 textbooks. Her research articles have appeared in journals of language and linguistics, such as *Studies of the Chinese Language*, *Studies in Language and Linguistics*, *Minority Languages of China* and *Chinese Linguistics*.