Reading in a Foreign Language October 2021, Volume 33, No. 2
ISSN 1539-0578 pp. 212–237

Constructing Diagnostic Reading Assessment Instruments for Low-level Chinese as Second Language Learners

Shuyi Yang
Johns Hopkins University
United States

Abstract

The present study investigated the applicability of word reading, word segmentation, and text reading as diagnostic tools to assess comprehension, detect struggling readers, and inform instruction for low-level Chinese as second language (L2) learners (*n* = 70). The results showed that the three instruments measured different dimensions of word recognition and predicted text comprehension. Text reading was the most robust indicator of comprehension and the most sensitive screener for weak readers. The diagnostic instruments provided interpretable feedback, located problems at specific areas, and evaluated instructional material difficulty level. The findings offer empirical support for the use of text reading for diagnostic purposes in Chinese low-level L2 reading instruction and suggest the importance of helping students develop word recognition skills.

***Keywords**:* diagnostic assessment, word reading, word segmentation, text reading, Chinese L2 reading comprehension

Reading in a second language (L2) is a complex process consisting of lower-level linguistic construction and higher-level semantic integration (Kintsch, 1988). Word recognition, a critical lower-level process (Grabe, 2009), poses heavy burdens for L2 readers, especially those at the initial phase of learning a novel orthography. The challenge is particularly formidable for Chinese L2 readers with an alphabetic first language (L1). Chinese (simplified script) has a logographic orthography with no grapheme-phoneme correspondence or demarcated word boundaries and thus unique word recognition processes. Given the enormous difficulty Chinese L2 learners face, it is vital to detect reader problems early and precisely. Compared with standardized tests that provide little diagnostic information, diagnostic assessment enables teachers to identify struggling readers, pinpoint their weaknesses, and initiate timely instructional accommodation (Gillet et al., 2011).

Despite its key role, diagnostic assessment for L2 reading is scarce. Most available instruments are designed for L1 readers (Nilsson, 2008, 2013), and the few L2 diagnostic tools almost exclusively target children learning an alphabetic L2 (Alderson et al., 2014). Moreover, given the linguistic properties of Chinese, it is essential to develop the specialized diagnostic assessment. The present study examines the applicability of word reading, word segmentation, and text reading as diagnostic instruments to measure comprehension, discriminate weak readers, and provide useful information for Chinese L2 low-proficiency learners.

Literature Review

Word recognition and comprehension

Word recognition, a process of mapping the orthographic form onto a representation in the mental lexicon, operates both at the lexical level with context-free word recognition, and the text level with understanding words embedded within connected texts. Context-free and in-context word recognition differs in demand from higher-level supra-lexical processing such as syntactic parsing and semantic integration (Kim, 2015; Rasinski et al., 2011; Stafura & Perfetti, 2017).

Accuracy and rate are two building blocks of word recognition (Kuhn et al., 2010; Morris et al., 2018). Efficient word recognition manifests in direct retrieval of correct words without conscious attention (Pasquarella et al., 2015). It is driven by high-quality lexical representations with orthographic, phonological, and semantic information tightly coupled and available synchronously.

Word recognition is the cornerstone of reading comprehension. It entails the understanding of individual word meaning that is continuously incorporated into a coherent textual representation. Inaccurate word recognition undermines sufficient meaning-extraction and meaning-construction. Moreover, according to Automaticity Theory (DeKeyser, 2001; LaBerge & Samuels, 1974; Segalowitz, 2003; Segalowitz & Segalowitz, 1993), lower- and higher-level processes compete for the readers' limited cognitive capacity. If massive cognitive resources are expended on lower-level processes, few will be available to engage in the higher-level processes underpinning comprehension. When lower-level processes are performed rapidly and non-deliberately, the overall cognitive load is reduced, and more attention will be allotted to higher-order processes. Word recognition is an essential lower-level process, and its automatic execution enables ample cognitive resources to support the higher-order processes that ultimately lead to comprehension. Therefore, word recognition is considered a prerequisite for comprehension. Stumbles during word recognition are the most common deficits among early struggling readers (Bonifacci & Tobia, 2016; Brasseur-Hock et al., 2011; Cho et al., 2019; Cirino et al., 2013; Clemens, Hsiao et al., 2020; Clemens, Simmons et al., 2017; Mancilla-Martinez et al., 2020). When word recognition becomes increasingly automatized, its impact diminishes. It is well documented that the contribution of word recognition to reading comprehension declines in the upper grades, whereas other skills such as oral language comprehension play a relatively more critical role (Foorman et al., 2018; Lee & Chen, 2019; Quellette & Beers, 2010; Verhoeven & van Leeuwe, 2012).

Chinese-specific word recognition processes

The way orthography encodes semantics and phonology varies across languages, influencing specific word recognition processes (Ziegler et al., 2010).

As a logographic language, Chinese lacks grapheme-phoneme correspondence (H. Shen, 2013). Each Chinese character corresponds to a syllable-level morpheme, not a phoneme (Perfetti, 2003). Phonemic decoding, namely, assembling phonemes represented in graphemes into the whole unit's pronunciation, is tremendously difficult in Chinese. Phonetic radicals are the only component that offers phonological clues, although these clues are unreliable and limited (Shu et al., 2003). Sounding out words without comprehension, a phenomenon common in alphabetic languages, is scarce in Chinese. Instead, the dominant pathway to obtain pronunciation is semantically mediated: Meaning is accessed before pronunciation is activated. The accurate articulation of a word relies on its successful identification.

Another unique feature of Chinese is the invisible word boundaries, which necessitates an additional word segmentation process. Chinese texts consist of continuous character strings, with individual words spatially unmarked. Specifying a word's beginning and end is an extra task beyond recognizing it in isolation. Readers process characters sequentially and evaluate the likelihood of adjacent characters comprising a word (Li & Pollatsek, 2020). When an individual word is segmented from surrounding characters, it is simultaneously recognized. In this regard, word segmentation and word recognition are unified as one process. Under other circumstances, segmentation is initiated to determine the contextually appropriate word out of multiple possible candidates. When ambiguity occurs, that is, a character can combine either with its left or right neighbouring character to form different words, correct segmentation is essential for deriving contextually appropriate meaning (H. Shen, 2018). Hence, word segmentation is a fundamental process in reading Chinese. The line of research on the effect of intentionally inserted word boundaries for L2 learners has generated positive results: Segmenting cues have been found to facilitate comprehension (Bai et al., 2010; Bassetti & Lu, 2016; Gao & Jiang, 2015; D. Shen et al., 2012; Yao, 2011; Zhou et al., 2020), highlighting the critical importance of word segmentation.

Diagnostic reading assessments

Diagnostic assessment, as its name indicates, aims at identifying weak readers and pinpointing their problems. It supplements standardized tests to demonstrate learner performance on various componential skills (Alderson et al., 2014). Such information allows for early detection and targeted remediation of learner difficulties in specific areas. Diagnostic assessment that is easy to administer and score will be readily applicable in pedagogical settings.

Read-aloud is a common diagnostic technique for word recognition. Examinees read aloud words either in lists or in connected texts, and their performance indicates word recognition competence. The accurate and rapid articulation builds upon phonological information activation, which is triggered by successful word recognition (Kearns & Ghanem, 2019; Nation & Cocksey, 2009). It should be noted that in alphabetic languages, correct word pronunciation can be achieved by sounding out phonemes without lexical access. However, this "decoding without comprehension" process works only for words that strictly follow grapheme-phoneme correspondence rules (McKay et al., 2008).

Word reading and text reading have been extensively used as effective screeners for L1 children with reading difficulties (Ball & O'Connor, 2016; Clemens, Shapiro, et al., 2011; Compton et al., 2006; Fuchs et al., 2004; Hasbrouck & Tindal, 2006; Hintze & Silberglitt, 2005; Johnson et al., 2010; Klingbeil et al., 2015; Parker et al., 2015; Shapiro et al., 2006; Stage & Jacobsen, 2001). A robust relationship has been established between word and text reading performance and comprehension measured by standardized tests (Hosp & Fuchs, 2005; Jenkins et al., 2003; McGlinchey & Hixson, 2004; Riedel, 2007; Roehrig et al., 2008; Sabatini et al., 2019; Schilling et al., 2007; Speece et al., 2010; Silberglitt et al., 2006; Torgesen et al., 2001; Wanzek et al., 2010; Wise et al., 2010; Wood, 2006), indicating strong criterion validity. Despite smaller correlation coefficients, the pattern holds among L2 learners (Crosson & Lesaux, 2010; Jeon, 2012; Jiang, 2016; Jiang et al., 2012; Lems, 2003; McTague et al., 2012).

The Informal Reading Inventory (IRI) is a systematically organized set of diagnostic instruments containing word reading, text reading, and comprehension questions (Gillet et al., 2011). IRIs have been developed for various L1s (Nilsson, 2008, 2013; Roe & Burns, 2010; Shanker & Cockrum, 2013) and have been cross-validated to standardized tests (Fountas & Pinnell, 2007). IRIs specialized for L2 readers thus far are scarce. The few existing L2 IRIs eliminate read-aloud tasks and keep only the comprehension question section (Lombardo, 1979; Stark, 1981), making them look more like standardized tests than diagnostic assessments.

The key function of diagnostic assessment is to diagnose reading problems. Diagnostic effectiveness, namely the accuracy in identifying readers who are or are not at risk, constitutes a crucial aspect of validity. To obtain diagnostic effectiveness, scholars first set a cut-off score for “fail” and “pass” status, usually the 25th percentile (Parker et al., 2015; Tong et al., 2015), and group students accordingly. They examine how well students’ performance on diagnostic assessment predicts their “fail or pass” membership. Research on diagnostic effectiveness has focused on text reading. The overall correct classification ranged between 66% and 80% (Ball & O’Connor, 2016; Hintze & Silberglitt, 2005; Johnson et al., 2010; Klingbeil et al., 2015; Parker et al., 2015; Shapiro et al., 2006; Stage & Jacobsen, 2001), suggesting adequate accuracy. The diagnostic effectiveness of word reading and IRIs has rarely been reported. The results of a handful of studies show that word reading has reasonable diagnostic effectiveness (Clemens, Shapiro, et al., 2011), whereas IRIs classify a far lower percentage of examinees successfully (Hazen, 2018; Parker et al., 2015).

As previously stated, the value of diagnostic assessment lies in its direct relevance to pedagogical practices. The primary purpose is to provide feedback to learners on the possible gap between their current level and the target level. The target performance is generally represented in three levels, following the IRI convention (Gillet et al., 2011). The *independent level* is characterized by fluent read-aloud and excellent comprehension. Students at the *instructional level* perform acceptably well, yet they still need assistance or instruction, as materials are challenging. The *frustration level* describes a weak reading with frequent errors and minimal comprehension. The scale for differentiating the levels comprises bands of percentages of words read aloud accurately and comprehension questions answered correctly. Efforts have also been undertaken to establish reading rate criteria (Hasbrouck & Tindal, 2006; Morris et al., 2018; Rasinski, 1999; Sabatini et al., 2019; Snow et al., 2018). Examinees compare their results against the scale to locate their position across the competence hierarchy and better understand their componential reading abilities. For Chinese reading, the criteria that define and distinguish performance levels are underdeveloped. No guidelines for level specifications or descriptors are available. Participant data in research on struggling readers has provided vague benchmarks for the frustration and instructional levels. Word reading accuracy for young readers who were identified as at-risk and likely read at the frustration level ranged from 45% to 55%, whereas average readers sustained about 70% accuracy, roughly equivalent to the instructional level (Fong & Ho, 2019; Tong et al., 2015; Zhang et al., 2014). Studies on Chinese children’s text reading are even rarer. Xue and colleagues (Xue et al., 2013) listed the mean sentence reading rate for second, fourth, and sixth graders, displaying stable growth from 150 characters per minute (CPM) to about 400 CPM. X. Wu and Anderson (2007) probed the text reading of 18 second graders and calculated the mean accuracy and rate for high, average, and low groups. The average and high groups performed similarly, with above 95% accuracy and above 120 CPM. The low group achieved similar high accuracy as their more skilled counterparts while lagging in reading rate.

Another goal of diagnostic assessment is to pinpoint the source of reading problems. Weak readers do not necessarily experience difficulties in the same component skills; instead, they exhibit high heterogeneity. By analyzing poor comprehenders' performance on a battery of diagnostic instruments, educators can track down the specific areas that require intervention. A prevalence of word recognition deficits, reflected in slow and inaccurate word and text reading, was observed among below-average L1 comprehenders (Bonifacci & Tobia, 2016; Brasseur-Hock et al., 2011; Cho et al., 2019; Cirino et al., 2013; Clemens, Hsiao et al., 2020; Clemens, Simmons et al., 2017; Zhang et al., 2014). In L2 reading, a prominent reader profile has emerged of low comprehension coupled with fair word naming (Bonifacci & Tobia, 2016; Bowyer-Crane et al., 2017; Cho et al., 2019; Lesaux & Harris, 2017; Sparks, 2015; Sparks & Luebbers, 2018). It is noticeable that these competent "word callers" were learners of an alphabetic L2. The grapheme-phoneme correspondence allows for direct print-pronunciation mapping without semantic involvement. L2 readers' advantage in sounding out consistent words and pseudo-words with exact grapheme-phoneme correspondence was in striking contrast to their severely limited vocabulary (Bonifacci & Tobia, 2016; Cho et al., 2019; Lesaux & Harris, 2017; Sparks, 2015; Sparks & Luebbers, 2018). It is reasonable to assume a more evident weakness in irregular word reading, supported solely by vocabulary, among less-skilled L2 comprehenders.

Finally, diagnostic assessment, specifically text reading, can evaluate the appropriateness of instructional materials. Scholars have found that text reading performance is closely associated with text difficulty level (Barth et al., 2014; Morris et al., 2018), suggesting the possibility of the former being an indicator of the latter. Nevertheless, the effectiveness of text reading in determining whether materials and learner ability are well matched remains unclear because of the paucity of research (Burns et al., 2015).

Regarding Chinese L2 reading, there is a dearth of diagnostic tools. Lü (2016) conducted word and text reading for young heritage learners, but she did not examine the validity of the instruments. In their study of beginners, H. Shen and Jiang (2013) reported a strong relationship between character reading and comprehension. This finding was replicated in research by Q. Wu et al. (2017), in which adequate accuracy in character reading was a distinguishing feature between low and intermediate proficiency levels. H. Shen and Jiang (2013) also developed a word segmentation test and observed a moderate correlation with comprehension. H. Shen (2019) administered a text reading measure to L2 learners across four proficiency levels. The correlation between read-aloud performance and comprehension decreased as learners' proficiency improved and became insignificant at the most advanced level. These results indicate that comprehension shifts from word-by-word recognition to textual integration increasingly tied to higher-order processes as learners progress in their L2 competence. Another contribution of this study was constructing text reading fluency criteria for the independent, instructional, and frustration levels.

The literature review has revealed some gaps in the Chinese L2 reading assessment. First, thus far, diagnostic tools designed for Chinese L2 readers are lacking. The few available L2 diagnostic instruments almost exclusively focus on learners of alphabetic L2s. It remains unclear whether these instruments are suitable for logographic Chinese. Chinese-specific reading processes, such as word segmentation, also require appropriate assessment.

Moreover, previous research on L2 diagnostic techniques has seldom explored the diagnostic effectiveness and provided sufficient diagnostic information. This study aims to fill the gaps by addressing the following research questions:

1. What is the relationship between low-level Chinese L2 learners' performance on the three instruments (word reading, word segmentation, text reading) and scores on a standardized reading test?
2. Can the three instruments accurately identify weak and strong L2 readers as determined by a standardized reading test?
3. What diagnostic information can the results of the three instruments provide?

Method

Participants

Seventy low-level Chinese L2 learners from four universities in the United States participated in this study (39 males and 31 females). Among them, 54 were English native speakers, and the rest were from alphabetic European language backgrounds. The preliminary sampling was based on institutional status; learners who had studied Chinese for two years at the post-secondary level were recruited. A standardized reading test was administered to determine participants' proficiency levels.

Instruments

The reading section of a College Board Advanced Placement Chinese Language and Culture Exam (hereafter, AP test) was used as a standardized test to determine participants' proficiency levels and to measure reading comprehension. The AP test has been widely utilized in the United States as a designated proficiency assessment for the two-year foreign language requirement at the college level. The test contains six texts of various genres, ranging from 133 to 382 characters in length, accompanied by 25 multiple-choice comprehension questions that measure textual information, inferences, and text purposes.

Word reading evaluated context-free word recognition. The researcher randomly selected a total of 25 words from the six texts in the AP test (see Appendix A) and ran Chinese TA Software to compute the word levels specified in the *Vocabulary and Character Proficiency Guideline* (National Office of Teaching Chinese as a Second Language, 1992). Among the 25 words, there were 11 Level 1 words, nine Level 2 words, two Level 3 words, one Level 4 word, and two words outside the *Guideline* list. The distribution of the words across levels was consistent with that in the 1,500 most frequent words.

The word segmentation test measured segmentation, a Chinese-specific dimension of in-context word recognition. The researcher randomly selected one sentence from each of the six texts in the AP test (see Appendix B) and checked the sentence difficulty level in terms of character and word frequency and sentence and sub-sentence length. The researcher collected the character and word frequency rank from *Modern Chinese Frequency Dictionary* (Beijing Language Institute Press, 1986). The word rank was around 700 to 1,500 (M = 1138.88), and the character rank was around 100 to 400 (M = 208.8), indicating that the sentences comprised high-frequency words and familiar characters. The sentence length ranged from 38 to 49 characters (M = 42.60). The sub-sentence refers to the sub-structure within a complex sentence, separated by punctuation marks. The mean sub-sentence length was about seven to 14 characters. In summary, sentences in the word segmentation test were comparable in difficulty level.

The text reading test assessed in-context word recognition. The researcher randomly retrieved one paragraph from each of the six texts in the AP test (see Appendix C), with length ranging from 81 to 190 characters (M = 130).

The researcher conducted the textbook-equivalent material tasks to investigate the usage of the three diagnostic instruments to estimate instructional material difficulty levels. The researcher chose *Integrated Chinese Level 2* (Liu et al., 2009) as the target textbook, which most participants in this study were currently using. The researcher randomly selected three texts from the textbook's beginning, middle, and end (L3, L9, L17). The three texts were 192, 164, and 239 characters long, respectively. Because all participants had finished the textbook by data collection, the researcher adopted alternative texts to ensure that participants had not learned them. The researcher retrieved another three texts (M1, M2, M3) from the beginning-level text corpus of *Chinese Reading World* (http://collections.uiowa.edu/chinese/) as the textbook-equivalent materials, with lengths of 206, 246, and 269 characters, respectively. The difficulty levels of the two sets of texts were matched in terms of character ranks and sentence length (see Table 1).

Table 1

Character Distribution and Average Sentence Length of Textbook and Matched Texts

			HSK Rank									
Text	CO	DC	Level 1		Level 2		Level 3		Level 4		Others	
			#	%	#	%	#	%	#	%	#	%
L3	192	101	94	93.07	4	3.96	2	1.98	0	0.00	1	0.99
M1	206	92	87	94.57	5	5.43	0	0.00	0	0.00	0	0.00
L9	164	90	80	88.89	9	10.00	0	0.00	0	0.00	1	1.11
M2	246	91	81	89.01	7	7.69	2	2.20	1	1.10	0	0.00
L17	239	145	116	80.00	26	17.93	3	2.07	0	0.00	0	0.00
M3	265	121	100	82.64	17	14.05	3	2.48	1	1.65	0	0.00

Text	Average Sentence Length
L3	21.33
M1	20.18
L9	23.43
M2	22.36
L17	30.13
M3	29.33

Note. CO = Character occurrences; DC = Distinct characters (a character that did not appear in the text before).

Chi-square analyses revealed that the matched texts were similar to the textbook texts in character rank: χ^2 (2) = 2.05, *p* = .36 for L3 and M1; χ^2 (3) = 3.23, *p* = .36 for L9 and M2; χ^2 (3) = 1.92, *p* = .59 for L17 and M3.

There was no significant difference between the two sets of texts in sentence lengths: *t* (18) = .40, *p* = .69 for L3 and M1; *t* (16) = .27, *p* = .79 for L9 and M2; *t* (15) = .17, *p* = .87 for L17 and M3. Therefore, the matched texts were comparable to the textbook texts in difficulty level.

The researcher developed three diagnostic tasks for the three matched texts. The researcher randomly selected twenty-five words for word reading, among which were 11 Level 1 words, nine Level 2 words, two Level 3 words, one Level 4 word, and two words outside the *Guideline* list. The researcher randomly chose one sentence from each of the three texts for the word segmentation test. The mean word frequency was around 700 to 1,500, and the mean character frequency was around 100 to 400. The sentence length ranged from 43 to 44 characters (*M* = 43.60), and the mean sub-sentence length was about eight to 14 characters. For text reading, the researcher randomly retrieved one paragraph from each of the three texts, ranging from 98 to 140 characters in length (*M* = 118).

Data collection

Participants performed all the tasks individually. During the 40-minute AP test, participants read the texts silently and answered the comprehension questions. For word reading, they read the words aloud one by one. On average, it took five minutes to finish this task. The word segmentation test, in which the participants marked word boundaries, lasted about five minutes. Text reading, in which the participants read the paragraphs aloud at a natural pace, generally lasted around 15 minutes.

There were two phases of data collection. In the first phase, the participants took the AP test and completed the three diagnostic assessments. They were randomly assigned to one of two groups. One group (*n* = 35) first took the AP test and then the three diagnostic assessments. The other group (*n* = 35) followed the reverse order. The procedure was counterbalanced to control the practice effect resulting from using the same materials twice in the AP test and in the three diagnostic instruments. In the second phase, participants completed the textbook-equivalent material tasks. The entire process lasted 90 minutes.

Scoring

AP test. Each correct answer was assigned one point. The accuracy score was the percentage of correctly answered items.

Word reading. Incorrect responses, unintelligible pronunciations, and skipped or indicated as unknown words were considered errors. The accuracy score was the percentage of correctly read words. The rate score was the number of characters read per minute (CPM). A second rater was recruited to code 20% of the data (15 participants). The inter-rater reliability coefficient was 0.99 (*p* < .01). The few disagreed-upon instances of scoring were negotiated and resolved.

Word segmentation test. Placing a boundary inside a word and combining non-word character strings as words were counted as errors. The accuracy score was the percentage of correct segmentations. A second rater coded 20% of the data. The inter-rater reliability coefficient was 0.80 (*p* < .01). A consensus on the disagreed-upon instances was reached by negotiation.

Text reading. Skips, substitutions, and hesitation for more than three seconds were treated as errors. Self-corrections, inappropriate pauses within words, and acceptable pronunciation deviations based on native speakers' judgment were not counted as errors. The accuracy score was the percentage of correctly read characters. The rate score was the number of characters read per minute (CPM). A second rater scored 20% of the data. The inter-rater reliability coefficient was 0.98 (*p* < .01). The few discrepancies in coding were negotiated and resolved.

Results

RQ 1: What is the relationship between learner performance on the three diagnostic instruments and scores on a standardized reading test?

The descriptive statistics are displayed in Table 2.

Table 2

Descriptive Statistics (n = 70)

	M	*SD*	Min	Max	Cronbach's Alpha
AP Chinese Test Accuracy (%)	56.51	20.43	16.00	96.00	0.83
Word Reading Accuracy (%)	59.43	21.72	4.00	100.00	0.91
Word Reading Rate (CPM)	31.50	21.18	2.22	103.84	
Word Segmentation Accuracy (%)	86.27	12.88	33.33	100.00	0.72
Text Reading Accuracy (%)	80.99	15.11	10.17	98.87	0.93
Text Reading Rate (CPM)	68.70	27.13	23.33	137.25	

First, the AP test score confirmed that the participants were at a low proficiency level. The Cronbach's alpha values of the instruments ranged from 0.72 to 0.93, indicating reasonably high reliability.

The data were transformed in response to non-linearity problems. The accuracy scores were transformed with the empirical logit function. The rate scores were log-transformed. Zero-order correlations were computed among six variables: word reading accuracy, word reading rate, word segmentation accuracy, text reading accuracy, text reading rate, and reading comprehension (see Table 3).

Table 3

Correlation Matrix Among the Six Measures

	1. RC	2. WRA	3. WRR	4. WSA	5. TRA	6. TRR
1. RC	--	.76**	.62**	.56**	.80**	.69**
2. WRA		--	.83**	.58**	.93**	.84**
3. WRR			--	.52**	.84**	.87**
4. WSA				--	.69**	.50**
5. TRA					--	.90**
6. TRR						--

Note. RC = reading comprehension measured by AP test accuracy; WRA = word reading accuracy; WRR = word reading rate; WSA = word segmentation accuracy; TRA = text reading accuracy; TRR = text reading rate. * *p* < .05; ** *p* < .01.

Students' comprehension performance on the standardized AP test strongly correlated with measures of the text reading and word reading. The correlation between comprehension and word segmentation was slightly weaker, yet still close to the large benchmark (*r* = .60, Plonsky & Oswald, 2014). All correlations were statistically significant at the .01 level.

Notably, strong correlations between accuracy and rate variables were detected, suggesting the two can converge with one composite score. Subsequently, Z scores of accuracy and rate were calculated and averaged to produce a holistic measure. The Z scores were transformed with the empirical logit function. Table 4 presents the results of zero-order correlation analyses involving four measures. The correlation between text reading and reading comprehension was the highest, followed by the correlation between word reading and word segmentation.

The inter-correlations among the three diagnostic instruments were high. To identify the unique contribution of each instrument to comprehension, multiple regression analyses with the AP score as the dependent variable were conducted.

Table 4

Correlation Matrix Among the Four Measures

	1. RC	2. WR	3. WS	4. TR
1. RC	--	.72**	.56**	.77**
2. WR		--	.58**	.93**
3. WS			--	.66**
4. TR				--

Note. RC = reading comprehension measured by AP test accuracy; WR = word reading; WS = word segmentation; TR = text reading. * *p* < .05; ** *p* < .01.

As shown in Table 5, word reading and word segmentation were statistically significant when they were the only predictors in the model. However, the two measures became insignificant when text reading was entered into the equation. In contrast, text reading made an independent, significant contribution to comprehension (*p* = .01) beyond word reading and word segmentation.

Table 5

Summary of Regression Models with Word Reading and Word Segmentation as First-order Variables

Model	b	SE	β	T	sig	Df	R^2	Adj R^2	△R^2	△F	Sig. F change
1 WR	.62	.11	.60	5.93	.00	2, 67	.55	.54	.55**	40.81	.00
WS	.22	.10	.22	2.17	.03						
2 WR	.08	.23	.08	.35	.73	3, 66	.59	.58	.04**	7.16	.01
WS	.10	.11	.10	.96	.34						
TR	.65	.24	.63	2.68	.01						

Note. WR = word reading; WS = word segmentation; TR = text reading. * $p < .05$; ** $p < .01$.

The results suggest that text reading was the most robust predictor of comprehension. Word reading and word segmentation did not explain the unique variance once the effect of text reading was controlled.

RQ2: Can the three diagnostic instruments accurately identify weak and strong L2 readers as determined by a standardized reading test?

Participants were divided into two groups based on their AP test raw scores, with the 25th percentile as the cut-off score. Learners below the 25th percentile were classified as the weak group (*n* = 19) and those at or above the 25th percentile were classified as "pass" group (*n* = 51). The three predictors were word reading (averaged Z score of accuracy and rate), word segmentation (Z score), and text reading (averaged Z score of accuracy and rate).

Logistic regression analyses were implemented to predict which learners were identified as "pass" L2 readers with the default .5 probability cut-off value. The initial model used all three predictors and provided a statistically better fit over the null model, χ^2 (3) = 18.52, $p < .001$. The Hosmer-Lemeshow test also showed a good fit, χ^2 (8) = 4.79, *p* = .78. This model's proportion reduction in deviance over the null model was 22.7% (R^2_L = .227; Negelkerke R^2 = .337). Only text reading was statistically significant according to Wald's test (See Table 6). The analysis was rerun after removing the non-significant predictors (word reading and word segmentation). The reduced model did not significantly differ from the full model in terms of fit, χ^2 (2) = .10, *p* = .95, and similarly had a statistically significant improvement over the null model, χ^2 (1) = 18.42, $p < .001$. A summary of the final, reduced model is presented in Table 6. This model contributed a 22.6% proportion reduction in deviance over the null model (R^2_L = .226; Negelkerke R^2 = .336) and yielded a good fit based on the Hosmer-Lemeshow test, χ^2 (8) = 3.57, *p* = .89. The overall classification was 80%. Text reading had positive coefficients and odds ratios of > 1, indicating that the increased scores led to a higher probability of becoming "pass" readers.

Table 6

Logistic Regression Analyses for L2 Comprehension Predictors

Variable	Full model					Final model				
	Coefficient	*SE*	Wald	*p*	Odds ratio	Coefficient	*SE*	Wald	*p*	Odds ratio
Constant	−2.39	1.92	1.56	.21	.09	−2.20	.92	5.73	.02*	.11
TR	.06	.03	4.47	.04*	1.06	.06	.02	10.81	.001**	1.03
WR	.02	.06	.08	.78	1.02					
WS	.34	2.55	.02	.89	1.41					

Note. WR = word reading; WS = word segmentation; TR = text reading. * *p* < .05; ** *p* < .01.

RQ3: What diagnostic information can the results of the three diagnostic instruments provide?

For each diagnostic instrument, participants' raw scores were divided into three levels (See Table 7): below the 25th percentile (low), between the 25th and 74th percentiles (average), and at or above the 75th percentile (high).

Table 7

Group Performance on Three Diagnostic Instruments

	WRA (%)				WRR (CPM)			
	M	*SD*	Min	Max	*M*	*SD*	Min	Max
L (17)	21.18	11.38	4.00	36.00	10.64	4.42	2.22	15.72
A (35)	56.11	9.48	40.00	68.00	26.80	7.06	16.00	40.34
H (18)	85.33	8.79	72.00	100.00	60.38	18.40	40.72	103.84
	WSA (%)							
	M	*SD*	Min	Max				
L (17)	68.59	13.47	33.33	79.76				
A (35)	89.12	4.50	82.14	95.24				
H (18)	97.42	1.18	96.30	100.00				
	TRA (%)				TRR (CPM)			
	M	*SD*	Min	Max	*M*	*SD*	Min	Max
L (17)	60.10	16.21	10.17	75.99	39.58	8.16	23.33	48.85
A (34)	84.16	3.40	78.25	89.55	63.19	10.11	49.10	83.00
H (19)	94.01	3.20	89.83	98.87	104.61	18.86	84.79	137.25

Note. L = low; A = average; H = high; WRA = word reading accuracy; WRR = word reading rate; WSA = word segmentation accuracy; TRA = text reading accuracy; TRR = text reading rate.

Based on the group performance, a tentative scale for independent, instructional, and frustration levels of word reading, word segmentation, and text reading was established (See Table 8).

Table 8

Scale for Three Levels of Reading

Reading level	WRA (%)	WRR (CPM)	WSA (%)	TRA (%)	TRR (CPM)
Independent	70% and up	41 and up	96% and up	90% and up	85 and up
Instructional	40%–69%	16–40	80%–95%	75%–89%	50–84
Frustration	Below 40%	Below 16	Below 80%	Below 75%	Below 50

Note. WRA = word reading accuracy; WRR = word reading rate; WSA = word segmentation accuracy; TRA = text reading accuracy; TRR = text reading rate.

Students can compare their performance with the scale and interpret their results. For example, participant 64 performed at the instructional level on word segmentation and text reading, but at the frustration level on word reading. Such a profile demonstrated weakness in context-free word recognition and suggested its focused training.

The performance on the three diagnostic instruments of the 19 weak comprehenders (below 25th percentile on AP test score) was further analyzed to locate the source of comprehension difficulties (See Table 9).

Table 9

Profiles of Weak Comprehenders

P	WRA	WRR	WSA	TRA	TRR
P.36	FRUS	FRUS	FRUS	FRUS	FRUS
P.19	FRUS	FRUS	FRUS	FRUS	FRUS
P.65	FRUS	FRUS	FRUS	FRUS	FRUS
P.15	FRUS	FRUS	FRUS	FRUS	INST
P.63	FRUS	FRUS	FRUS	FRUS	FRUS
P.67	FRUS	FRUS	INST	FRUS	FRUS
P.70	FRUS	FRUS	INST	FRUS	FRUS
P.68	FRUS	FRUS	INST	FRUS	FRUS
P.62	FRUS	FRUS	FRUS	FRUS	INST
P.22	FRUS	FRUS	FRUS	FRUS	INST
P.11	INST	INST	INST	FRUS	INST
P.56	FRUS	FRUS	INST	FRUS	FRUS
P.18	INST	INST	FRUS	INST	INST
P.34	INST	IND	FRUS	INST	IND
P.35	INST	IND	INST	INST	INST
P.9	INST	INST	FRUS	INST	INST
P.55	IND	INST	INST	INST	INST
P.47	INST	INST	INST	INST	INST
P.28	INST	INST	IND	INST	INST

Note. P = Participant No; IND = independent level; INST = instructional level; FRUS = frustration level.

Among the 19 weak comprehenders, four were at the frustration level on all five measures. Seven performed at the frustration level on four measures, while achieved average performance on word segmentation accuracy or text reading rate. Four showed weakness in only one measure (word segmentation accuracy or text reading accuracy) and did well on other measures. The other four students reached the instructional or independent level on all measures, suggesting that their comprehension difficulties cannot be attributed to deficient word recognition but other reading skills.

Another application of the scale was to evaluate the difficulty level of instructional materials. Since text reading functioned best in predicting comprehension and screening problematic readers, text reading accuracy and rate scores of the textbook-equivalent materials were used for comparison against the scale. When the estimated reading levels differed for the accuracy and rate benchmarks, the lower level was taken. Table 10 displays the percentage of participants at each of the three levels for the three texts.

Table 10

Percentage of Estimated Level Based on Text Reading

Estimated Level	Text		
	1	2	3
IND	32.90%	34.30%	11.40%
INST	47.10%	47.10%	32.90%
FRUS	20.00%	18.60%	55.70%

Note. IND = independent level; INST = instructional level; FRUS = frustration level.

Texts 1 and 2 are at the instructional level for 47.10% of the participants and the independent level for about 30%. Only about 20% of the participants fall in the frustration level for the two texts. By contrast, text 3 is at the frustration level for more than half of the participants. The results indicated that texts 1 and 2 are appropriate for the students as instructional materials, whereas text 3 is too difficult and should not be used for classroom teaching.

Discussion

Effectiveness of three diagnostic instruments to assess comprehension

One major finding of the analyses was that all three diagnostic instruments effectively assessed comprehension, indicating robust criterion validity. Text reading emerged as the most potent predictor, in line with previous studies on both L1 (McGlinchey & Hixson, 2004; Riedel, 2007; Roehrig et al., 2008; Schilling et al., 2007; Silberglitt et al., 2006; Torgesen et al., 2001; Wanzek et al., 2010; Wood, 2006) and L2 readers (Crosson & Lesaux, 2010; Jeon, 2012; Jiang, 2016; Jiang et al., 2012; Lems, 2003; McTague et al., 2012; H. Shen, 2019). Word reading also contributed substantially to comprehension, adding to the mounting evidence in L1 (Hosp & Fuchs, 2005; Jenkins et al., 2003; Sabatini et al., 2019; Speece et al., 2010; Wise et al., 2010) and L2 research (Crosson & Lesaux, 2010; Jeon, 2012; Jiang et al., 2012; H. Shen & Jiang, 2013; Q. Wu et al., 2017).

It is noteworthy that the magnitude of correlations involving the two read-aloud tasks was comparable to that in research on Chinese L2 beginners (H. Shen & Jiang, 2013; H. Shen, 2019), but much stronger than what was reported for both EFL readers (Crosson & Lesaux, 2010; Jeon, 2012; Jiang et al., 2012; Lems, 2003; McTague et al., 2012) and advanced Chinese learners (H. Shen, 2019; Q. Wu et al., 2017). The discrepancy between the present study and the EFL research can be attributed to orthography differences. In alphabetic English, readers can bind together phonemes into word pronunciation without knowing its meaning, as long as they have acquired the grapheme-phoneme correspondence rules. The orthography-phonology route does not entail the activation of semantic information and, thus, oral reading realized via this route is less tied to comprehension. By contrast, logographic Chinese lacks the phoneme-level script-sound correspondence, rendering read-aloud a semantic processing task. It is necessary to access word meaning to derive pronunciation. In other words, reading aloud in Chinese, to a more significant extent, shares semantically oriented mechanisms with comprehension. Hence, the two have a stronger relationship.

As for the divergence from studies on advanced Chinese learners, there were two possible explanations. First, the relatively lower correlations among high-proficiency L2 learners, as compared with the present study, demonstrated the downward trend in the association between word recognition and comprehension as reading ability grows (Foorman et al., 2018; Quellette & Beers, 2010; Sabatini et al., 2019; Verhoeven & van Leeuwe, 2012). There is a wide consensus that once word recognition reaches a threshold, it no longer acts as a central factor in reading comprehension. Another interpretation pertains to the features of reading materials. The words and texts in this study are relatively simple and do not place high demands on vocabulary knowledge. It is attested in the literature that L2 oral reading performance and comprehension are highly correlated for easy materials and that correlation strength weakens for authentic, complex texts (Crossley et al., 2014; Lee & Chen, 2019). Moreover, materials for advanced learners tend to have denser content and more sophisticated structures that require high-order processes, such as inference generation and information synthesis beyond understanding individual word meaning (H. Shen, 2019).

The correlation between word segmentation and comprehension was moderate, albeit significant, congruent with H. Shen and Jiang (2013). As a uniquely invoked process of contextual word recognition in Chinese, word segmentation was directly linked to comprehension, consistent with the observed benefits of the visually marked word boundaries for L2 readers (Bai et al., 2010; Bassetti & Lu, 2016; Gao & Jiang, 2015; D. Shen et al., 2012; Yao, 2011; Zhou et al., 2020). Efficient word segmentation packs continuous characters into meaningful words, thereby maintaining semantic coherence. The failure to detect the word boundary jeopardizes the identification of embedded words and further disturbs meaning construction.

The high inter-correlations among the three diagnostic instruments provided evidence for construct validity. The three diagnostic instruments tapped different aspects of one construct: word recognition. On the one hand, the three instruments were correlated highly enough to be theoretically meaningful, indicating high convergent validity. On the other hand, the three instruments remained separate sub-constructs and thus achieved reasonable discriminant validity.

Word recognition in Chinese encompasses three dimensions: accuracy, rate, and segmentation. Accuracy and rate are two primary indices of word recognition (Kuhn et al., 2010; Morris et al., 2018). Proficient readers rapidly and effortlessly access a given word and extract its complete information, reading accurately, smoothly, and appropriately. Weak readers, due to their sparse stored lexical representations and insufficient word processing skills, struggle with analyzing the orthographic form and activating the target word. Therefore, their reading is prone to errors, halting, and filled with pauses. Distinguished from alphabetic languages, Chinese in-context word recognition has an additional component of segmentation. The absence of word boundaries invokes a process of dividing character sequences into meaningful words, co-occurring with word identification (Li & Pollatsek, 2020). Strong parsers automatically and precisely detect embedded words and separate them. In contrast, less-skilled parsers hesitate at judging which characters form a word and tend to segment in a way that yields non-words and disrupts meaning.

The significant correlations between the three diagnostic instruments and reading comprehension highlighted the crucial role of word recognition. As cognitive competition theory claims (DeKeyser, 2001; LaBerge & Samuels, 1974; Segalowitz, 2003; Segalowitz & Segalowitz, 1993), cognitive resources are limited. Laborious and effortful word recognition occupies a disproportionate amount of attention, with relatively little attention devoted to higher-order processes. Consequently, comprehension is hindered. As L2 readers gain a better command of word recognition skills, sufficient cognitive resources are freed up for higher-level processes to take place, which in turn promotes comprehension.

Finally, it was noted that text reading uniquely predicted comprehension over and above word reading and word segmentation. Reading connected texts relies on the orchestration of multiple skills, including, but not limited to, word recognition. Text reading involves analyzing syntactic structures and incorporating semantic elements in sentences (Kim, 2015; Rasinski et al., 2011; Stafura & Perfetti, 2017), and thus captures some comprehension processes in addition to word recognition.

Effectiveness of three diagnostic instruments to identify weak L2 readers

In this study, text reading displayed adequate diagnostic effectiveness in discriminating weak L2 readers. The overall correct classification reached 80%, similar to L1 studies (Ball & O’Connor, 2016; Hintze & Silberglitt, 2005; Johnson et al., 2010; Klingbeil et al., 2015; Parker et al., 2015; Shapiro et al., 2006; Stage & Jacobsen, 2001). L2 educators who use text reading indices would be correct most of the time in deciding whether a student reaches or fails to reach the criterion. It is encouraging that text reading had a good capacity to distinguish poor from competent L2 readers and was as accurate as identifying at-risk L1 child readers.

Diagnostic information of three diagnostic instruments

The present study proposes tentative criteria for independent, instructional, and frustration levels of L2 Chinese reading in the service of score interpretation. For text reading, the accuracy range at the instructional level (75% to 89%) approximated the mean (80% to 88%) among Chinese heritage children (Lü, 2016). The entire rate scale overlapped mostly with H. Shen’s (2019), supporting its effectiveness in differentiating performance ranks. In terms of word reading, the instructional-level rate benchmark (16 to 40 CPM) was commensurate with the placement specification (28 to 40 CPM) for novice L2 foreign learners (Q. Wu et al., 2017), but below the average (57 CPM) of young heritage readers (Lü, 2016), reflecting a gap between heritage and foreign learners. Compared with L1 children, L2 adult beginners at the independent level approached the mean accuracy of L1 children in the early grades (Fong &

Ho, 2019; Tong et al., 2015; Zhang et al., 2014), but read at less than 60% of the average speed (Xue et al., 2013; X. Wu & Anderson, 2007). It converged with previous findings of L2 reading being close to L1 norm in accuracy while far behind in rate (Jeon, 2012; Jiang et al., 2012; Lems, 2003). Finally, the word segmentation accuracy minimum was reasonably high (80%), comparable to the ceiling level observed in H. Shen and Jiang (2013). The universal superior performance on word segmentation produced little variability in the data, explaining the moderate correlation with comprehension. It also indicated that word segmentation functioned less well in discriminating different levels.

The current study constructed profiles of struggling comprehenders and explored the source of their problems. The majority fell in the word stumbler category, as they exhibited deficits in one or multiple aspects of word recognition. The pervasive weakness in word recognition has been well documented in L1 at-risk readers (Bonifacci & Tobia, 2016; Brasseur-Hock et al., 2011; Cho et al., 2019; Cirino et al., 2013; Clemens, Hsiao et al., 2020; Clemens, Simmons et al., 2017; Zhang et al., 2014), reiterating the gatekeeper role of word recognition for comprehension. Four poor comprehenders performed at or above the average level on all measures, resembling the word caller profile (Bonifacci & Tobia, 2016; Bowyer-Crane et al., 2017; Cho et al., 2019; Lesaux & Harris, 2017; Sparks, 2015; Sparks & Luebbers, 2018). They encountered difficulties in processes other than word recognition. Notably, this type of reader accounted for a small proportion, probably because Chinese logographic orthography restricts the sounding out of words that underlies accurate yet semantically unengaged reading loud.

Regarding the evaluation of text difficulty, text reading proved to be a potentially helpful tool. Learner performance aligned well with text difficulty, consistent with previous studies (Barth et al., 2014; Morris et al., 2018). Reading aloud a text segment is a quick and straightforward way to judge whether the material is appropriate or whether a modification is needed.

Conclusion

The present study examined the validity and usability of three diagnostic instruments in Chinese low-level L2 reading. The results showed that all three instruments effectively assessed reading comprehension. Word recognition emerged as paramount in reading comprehension and should be placed at the core of L2 reading instruction. Pedagogical approaches such as reading aloud, repeated reading, and extensive reading should be incorporated throughout the curriculum.

Text reading was the strongest predictor of comprehension and the most precise screener of low comprehenders. It effectively distinguished L2 readers at different levels and estimated the instructional material difficulty. The evidence pointed to text reading as a good candidate for diagnostic assessment in Chinese low-level L2 reading. It especially fits Chinese because the lack of script-sound correspondence makes reading aloud a semantically oriented task and thus a more robust measure of comprehension-related processes. Furthermore, text reading can be performed with available materials and does not require extensive training to administer and score.

The current study demonstrated the diagnostic information the three instruments provided to inform instruction. A preliminary scale for independent, instructional, and frustration levels was established so that examinees could easily interpret their scores. The battery of multiple instruments allowed for the location of learner weaknesses and corresponding targeted remediation.

There are several limitations of this study. First, this research focused on novice L2 readers. It remains to be seen whether the findings can be generalized to advanced learners. A related issue is to track the dynamic interaction between word recognition and comprehension and proficiency development and, more importantly, to determine the threshold of word recognition ability for adequate comprehension. It would also be valuable to construct and validate diagnostic tools for other reading skills beyond word recognition. This study reported a few poor comprehenders with sufficient word recognition, which indicates the need to conduct further tests to identify the cause of their problems.

References

Alderson, J. C., Haapakangas, E.-L., Huhta, A., Nieminen, L., & Ullakonoja, R. (2014). *The diagnosis of reading in a second or foreign language*. Routledge.

Bai, X., Zhang, T., Tian, L., Liang, F., & Wang, T. (2010). Effect of word segmentation on American students reading Chinese: Evidence from eye movements. *Psychological Research*, *5*, 25−30.

Ball, C. R., & O'Connor, E. (2016). Predictive utility and classification accuracy of oral reading fluency and the measures of academic progress for the Wisconsin knowledge and concepts exam. *Assessment for Effective Intervention, 41*(4), 195−208. https://doi.org/10.1177/1534508415620107

Barth, A. E., Tolar, T. D., Fletcher, J. M., & Francis, D. (2014). The effects of student and text characteristics on the oral reading fluency of middle-grade students. *Journal of Educational Psychology, 106*(1), 162−180. https://doi.org/10.1037/a0033826

Bassetti, B., & Lu, M. (2016). Effects of interword spacing on native English readers of Chinese as a second language. *International Review of Applied Linguistics in Language Teaching, 54*(1), 1−22. http://dx.doi.org/10.1515/iral-2016-0014

Beijing Language Institute Press (1986). *Modern Chinese frequency dictionary*. Beijing Language Institute Press.

Bonifacci, P., & Tobia, V. (2016). Crossing barriers: Profiles of reading and comprehension skills in early and late bilinguals, poor comprehenders, reading impaired, and typically developing children. *Learning and Individual Differences, 47*, 17−26. https://doi.org/10.1016/j.lindif.2015.12.013

Bowyer-Crane, C., Fricke, S., Schaefer, B., Lervag, A., & Hulme, C. (2017). Early literacy and comprehension skills in children learning English as an additional language and monolingual children with language weaknesses. *Reading and Writing, 30*(4), 771−790. https://doi.org/10.1007/s11145-016-9699-8

Brasseur-Hock, I. F., Hock, M. F., Kieffer, M. J., Biancarosa, G., & Deshler, D. D. (2011). Adolescent struggling readers in urban schools: Results of a latent class analysis. *Learning and Individual Differences, 21*(4), 438−452. https://doi.org/10.1016/j.lindif.2011.01.008

Burns, M. K., Pulles, S. M., Maki, K. E., Kanive, R., Hodgson, J., Helman, L. A., McComas, J. J., & Preast, J. L. (2015). Accuracy of student performance while reading leveled books rated at their instructional level by a reading inventory. *Journal of School Psychology, 53*(6), 437−445. https://doi.org/10.1016/j.jsp.2015.09.003

Cho, E., Capin, P., Roberts, G., Roberts, G., & Vaughn, S. (2019). Examining sources and mechanisms of reading comprehension difficulties: Comparing English learners and non-English learners within the simple view of reading. *Journal of Educational Psychology, 111*(6), 982−1000. https://doi.org/10.1037/edu0000332

Cirino, P. T., Romain, M. A., Barth, A. E., Tolar, T. D., Fletcher, J. M., & Vaughn, S. (2013). Reading skill components and impairments in middle school struggling readers. *Reading and Writing, 26*(7), 1059–1086. https://doi.org/10.1007/s11145-012-9406-3

Clemens, N. H., Hsiao, Y., Lee, K., Martinez-Lincoln, A., Moore, C., Toste, J., & Simmons, L. (2020). The differential importance of component skills on reading comprehension test performance among struggling adolescent readers. *Journal of Learning Disabilities, 54*(3), 155–169. https://doi.org/10.1177/0022219420932139

Clemens, N. H., Shapiro, E. S., & Thoemmes, F. (2011). Improving the efficacy of first grade reading screening: An investigation of word identification fluency with other early literacy indicators. *School Psychology Quarterly, 26*(3), 231–244. https://doi.org/10.1037/a0025173

Clemens, N. H., Simmons, D., Simmons, L. E., Wang, H., & Kwok, O. (2017). The prevalence of reading fluency and vocabulary difficulties among adolescents struggling with reading comprehension. *Journal of Psychoeducational Assessment, 35*(8), 785–798. https://doi.org/10.1177/0734282916662120

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology, 98*(2), 394–409. https://doi.org/10.1037/0022-0663.98.2.394

Crossley, S. A., Yang, H. S., & McNamara, D. S. (2014). What's so simple about simplified texts? A computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language, 26*(1), 92–113.

Crosson, A. C., & Lesaux, N. K. (2010). Revisiting assumptions about the relationship of fluent reading to comprehension: Spanish-speakers' text reading fluency in English. *Reading and Writing, 23*(5), 475–494. https://doi.org/10.1007/s11145-009-9168-8

DeKeyser, R. M. (2001). Automaticity and automatization. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 125–151). Cambridge University Press.

Fong, C. Y., & Ho, C. S. (2019). Poor oral discourse skills are the key cognitive-linguistic weakness of Chinese poor comprehenders: A three-year longitudinal study. *First Language, 39*(3), 281–297. https://doi.org/10.1177/0142723719830868

Foorman, B., Petscher, Y., & Herrera, S. (2018). Unique and common effects of decoding and language factors in predicting reading comprehension in grades 1–10. *Learning and Individual Differences, 63*, 12–23. https://doi.org/10.1016/j.lindif.2018.02.011

Fountas, I. C., & Pinnell, G. S. (2007). *Benchmark assessment system*. Heinemann.

Fuchs, D., Fuchs, L. S., & Compton, D. L. (2004). Identifying reading disability by responsiveness-to-instruction: Specifying measures and criteria. *Learning Disability Quarterly, 27*(4), 216–222. https://doi.org/10.2307/1593674

Gao, S., & Jiang, X. (2015). The effect of word boundaries on Chinese reading for second language learners. *Language Teaching and Linguistics Studies, 4*, 5–17.

Gillet, J. W., Temple, C., Temple, C., & Crawford, A. (2011). *Understanding reading problems: Assessment and instruction* (8th ed.). Pearson.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge University Press.

Hasbrouck, J. E., & Tindal, G. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher, 59*(7), 636–644. https://doi.org/10.1598/RT.59.7.3

Hazen, E. A. (2018). *The relationship between screening measures for reading and performance on the end of year state assessment in third grade*. [Unpublished doctoral dissertation]. University of Houston.

Hintze, J. M., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R–CBM and high-stakes testing. *School Psychology Review, 34*(3), 372–386. https://doi.org/10.1080/02796015.2005.12086292

Hosp, M. K., & Fuchs, L. S. (2005). Using CBM as an indicator of decoding, word reading, and comprehension: Do the relations change with grade? *School Psychology Review, 34*(1), 9–26. https://doi.org/10.1080/02796015.2005.12086272

Jenkins, J. R., Fuchs, L. S., Van den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology, 95*(4), 719–729. https://doi.org/10.1037/0022-0663.95.4.719

Jeon, E. H. (2012). Oral reading fluency in second language reading. *Reading in a Foreign Language, 24*(2), 186–208.

Jiang, X. (2016). The role of oral reading fluency in ESL reading comprehension among learners of different first language backgrounds. *Reading Matrix: An International Online Journal, 16*(2), 227–242.

Jiang, X., Sawaki, Y., & Sabatini, J. (2012). Word reading efficiency, text reading fluency, and reading comprehension among Chinese learners of English. *Reading Psychology, 33*(4), 323–349. https://doi.org/10.1080/02702711.2010.526051

Johnson, E. S., Jenkins, J. R., & Petscher, Y. (2010). Improving the accuracy of a direct route screening process. *Assessment for Effective Intervention, 35*(3), 131–140. https://doi.org/10.1177/1534508409348375

Kearns, D. M., & Ghanem, R. A. (2019). The role of semantic information in children’s word reading: Does meaning affect readers’ ability to say polysyllabic words aloud? *Journal of Educational Psychology, 111*(6), 933–956. https://doi.org/10.1037/edu0000316

Kim, A. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32*(2), 227–258. https://doi.org/10.1177/0265532214558457

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction–integration model. *Psychological Review, 95*(2), 163–182. https://doi.org/10.1037/0033-295X.95.2.163

Klingbeil, D. A., McComas, J. J., Burns, M. K., & Helman, L. (2015). Comparison of predictive validity and diagnostic accuracy of screening measures of reading skills. *Psychology in the Schools, 52*(5), 500–514. https://doi.org/10.1002/pits.21839

Kuhn, M. R., Schwanenflugel, P. J., & Meisinger, E. B. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly, 45*(2), 230–251. https://doi.org/10.1598/RRQ.45.2.4

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*(2), 293–323. https://doi.org/10.1016/0010-0285(74)90015-2

Lee, K., & Chen, X. (2019). An emergent interaction between reading fluency and vocabulary in the prediction of reading comprehension among French immersion elementary students. *Reading and Writing, 32*(7), 1657–1679. https://doi.org/10.1007/s11145-018-9920-z

Lems, K. (2003). *Adult ESL oral reading fluency and silent reading comprehension*. [Unpublished doctoral dissertation]. National Louis University.

Lesaux, N. K., & Harris, J. R. (2017). An investigation of comprehension processes among adolescent English learners with reading difficulties. *Topics in Language Disorders, 37*(2), 182–203. https://doi.org/10.1097/TLD.0000000000000120

Li, X., & Pollatsek, A. (2020). An integrated model of word processing and eye-movement control during Chinese reading. *Psychological Review, 127*(6), 1139–1162. https://doi.org/10.1037/rev0000248

Liu, Y., Yao, T., Bi, N., Ge, L., & Shi, Y. (2009). *Integrated Chinese Level 2* (3rd ed.). Cheng & Tsui Company.

Lombardo, M. (1979). *The effectiveness of an informal reading inventory in identifying the functional reading levels of bilingual students*. (Bilingual Education Paper Series Vol. 2 No. 10). Evaluation, Dissemination and Assessment Center.

Lü, C. (2016). Chinese reading development among young learners in a Chinese immersion program and a Chinese heritage language school. *Chinese Teaching in the World, 30*, 550–562.

Mancilla-Martinez, Hwang, Oh, & McClain (2020). Early elementary grade dual language learners from Spanish-speaking homes struggling with English reading comprehension: The dormant role of language skills. *Journal of Educational Psychology, 112*(5), 880–894. https://doi.org/10.1037/edu0000402

McGlinchey, M. T., & Hixson, M. D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review, 33*(2), 193–203. https://doi.org/10.1080/02796015.2004.12086242

McKay, A., Davis, C., Savage, G., & Castles, A. (2008). Semantic involvement in reading aloud: Evidence from a non-word training study. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 34*(6), 1495–1517. https://doi.org/10.1037/a0013357

McTague, B., Lems, K., Butler, D., & Carmona, E. (2012). ELL fluency scores—What can they tell us? In T. Rasinski, C. Blachowicz, & K. Lems (Eds.), *Fluency instruction: Research-based best practices* (pp. 278–288). The Guilford Press.

Morris, D., Pennell, A., Perney, J., & Trathen, W. (2018). Using subjective and objective measures to predict level of reading fluency at the end of first grade. *Reading Psychology, 39*(3), 253–270. https://doi.org/10.1080/02702711.2017.1418466

Nation, K., & Cocksey, J. (2009). The relationship between knowing a word and reading it aloud in children's word reading development. *Journal of Experimental Child Psychology, 103*(3), 296–308. https://doi.org/10.1016/j.jecp.2009.03.004

National Office of Teaching Chinese as a Second Language (1992). *Vocabulary and character proficiency guidelines*. Beijing Language Institute Press.

Nilsson, N. L. (2008). A critical analysis of eight informal reading inventories. *The Reading Teacher, 61*(7), 526–536. https://doi.org/10.1598/RT.61.7.2

Nilsson, N. L. (2013). The reliability of informal reading inventories: What has changed? *Reading and Writing Quarterly, 29*(3), 208–230. https://doi.org/10.1080/10573569.2013.789779

Parker, D. C., Zaslofsky, A. F., Burns, M. K., Kanive, R., Hodgson, J., Scholin, S. E., & Klingbeil, D. A. (2015). A brief report of the diagnostic accuracy of oral reading fluency and reading inventory levels for reading failure risk among second- and third-grade students. *Reading and Writing Quarterly, 31*(1), 56–67. https://doi.org/10.1080/10573569.2013.857970

Pasquarella, A., Chen, X., Gottardo, A., & Geva, E. (2015). Cross-language transfer of word reading accuracy and word reading fluency in Spanish-English and Chinese–English bilinguals: Script-universal and script-specific processes. *Journal of Educational Psychology, 107*(1), 96–110. https://doi.org/10.1037/a0036966

Perfetti, C. A. (2003). The universal grammar of reading. *Scientific Studies of Reading, 7*(1), 3–21. https://doi.org/10.1207/S1532799XSSR0701_02

Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079

Quellette, G., & Beers, A. (2010). A not-so-simple view of reading: How oral vocabulary and visual-word recognition complicate the story. *Reading and Writing, 23*(2), 189–208. https://doi.org/10.1007/s11145-008-9159-1

Rasinski, T. V. (1999). Exploring a method for estimating independent, instructional, and frustration reading rates. *Journal of Reading Psychology, 20*(1), 61–69. https://doi.org/10.1080/027027199278501

Rasinski, T. V., Reutzel, D. R., Chard, D., & Linan-Thompson, S. (2011). Reading fluency. In M. L. Kamil, P. D. Pearson, E. Birr Moje, & P. P. Afflerbach (Eds.), *Handbook of reading research: Volume IV* (pp. 286–319). Routledge.

Riedel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban, first grade students. *Reading Research Quarterly, 42*(4), 460–466. https://doi.org/10.1598/RRQ.42.4.5

Roe, B., & Burns, P. C. (2010). *Informal reading inventory: Preprimer to twelfth grade* (8th ed.). Wadsworth.

Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Accuracy of the DIBELS oral reading fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology, 46*(3), 343–366. https://doi.org/10.1016/j.jsp.2007.06.006

Sabatini, J., Wang, Z., & O’Reilly, T. (2019). Relating reading comprehension to oral reading performance in the NAEP fourth-grade special study of oral reading. *Reading Research Quarterly, 54*(2), 253–271. https://doi.org/10.1002/rrq.226

Schilling, S. G., Carlisle, J. F., Scott, S. E., & Zeng, J. (2007). Are fluency measures accurate predictors of reading achievement? *The Elementary School Journal, 107*(5), 429–448.

Segalowitz, N. S. (2003). Automaticity and second language learning. In C. Doughty & M. Long (Eds.), *The handbook of second language acquisition* (pp. 382–408). Blackwell.

Segalowitz, N. S., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics, 14*(3), 369–385. https://doi.org/10.1017/S0142716400010845

Shanker, J. L., & Cockrum, W. (2013). *Ekwall/Shanker reading inventory* (6th ed.). Pearson.

Shapiro, E. S., Keller, M. A., Lutz, J. G., Santoro, L. A., & Hintze, J. M. (2006). Curriculum-based measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment, 24*(1), 19–35. https://doi.org/10.1177/0734282905285237

Shen, H. H. (2013). Chinese L2 literacy development: Cognitive characteristics, learning strategies, and pedagogical interventions. *Language and Linguistics Compass, 7*(7), 371–387. https://doi.org/10.1111/lnc3.12034

Shen, H. H. (2018). Chinese as a second language reading: Lexical access and text comprehension. In C. Ke (Ed.), *The Routledge handbook of Chinese second language acquisition* (pp. 134–150). Routledge.

Shen, H. H. (2019). An investigation on instructional-level reading among Chinese L2 learners. *Journal of Second Language Teaching and Research, 7*(1), 184–211.

Shen, H. H., & Jiang, X. (2013). Character reading fluency, word segmentation accuracy, and reading comprehension in L2 Chinese. *Reading in a Foreign Language, 25*(1), 1–25.

Shen, D., Liversedge, S. P., Tian, J., Zang, C., Cui, L, Bai, X., Yan, G., & Rayner, K. (2012). Eye movements of second language learners when reading spaced and unspaced

Chinese text. *Journal of Experimental Psychology: Human Perception and Performance, 18*(2), 192–202. https://doi.org/10.1037/a0027485

Shu, H., Chen, X., Anderson, R. C., Wu, N., & Xuan, Y. (2003). Properties of school Chinese: Implications for learning to read. *Child Development, 74*(1), 27–47. https://doi.org/10.1111/1467-8624.00519

Silberglitt, B., Burns, M. K., Madyun, N. H., & Lail, K. E. (2006). Relationship of reading fluency assessment data with state accountability test scores: A longitudinal comparison of grade levels. *Psychology in the Schools*, *43*(5), 527–535. https://doi.org/10.1002/pits.20175

Snow, A. B., Morris, D., & Perney, J. (2018). Evaluating the effectiveness of a state-mandated benchmark reading assessment: mClass Reading 3D (Text reading and comprehension). *Reading Psychology*, *39*(4), 303–334. https://doi.org/10.1080/02702711.2017.1422302

Sparks, R. L. (2015). Language deficits in poor L2 comprehenders: The simple view. *Foreign Language Annals*, *48*(4), 635–658. https://doi.org/10.1111/flan.12163

Sparks, R. L., & Luebbers, J. (2018). How many U.S. high school students have a foreign language reading “disability”? Reading without meaning and the simple view. *Journal of Learning Disabilities, 51*(2), 194–208. https://doi.org/10.1177/0022219417704168

Speece, D. L., Ritchey, K. D., Silverman, R., Schatschneider, C., Walker, C. Y., & Andrusik, K. N. (2010). Identifying children in middle childhood who are at risk for reading problems. *School Psychology Review*, *39*(2), 258–276.

Stafura, J., & Perfetti, C. (2017). Integrating word processing with text comprehension: Theoretical frameworks and empirical examples. In K. Cain, D. L. Compton, & R. K. Parilla (Eds.), *Theories of reading development* (pp. 9–32). John Benjamins.

Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review, 30*(3), 407–419. https://doi.org/10.1080/02796015.2001.12086123

Stark, M. W. (1981). *A group informal reading inventory: An instrument for the assessment of ESL students' reading performance*. [Unpublished doctoral dissertation]. Oregon State University.

The University of Iowa Chinese Program (n.d.). *Chinese Reading World*. http://collections.uiowa.edu/chinese/

Tong, X., Tong, X., & McBride-Chang, C. (2015). A tale of two writing systems: Double dissociation and meta-linguistic transfer between Chinese and English word reading among Hong Kong children. *Journal of Learning Disabilities, 48*(2), 130–145. https://doi.org/10.1177/0022219413492854

Torgesen, J. K., Rashotte, C. A., & Alexander, A. W. (2001). Principles of fluency instruction in reading: Relationships with established empirical outcomes. In M. Wolf (Ed.), *Dyslexia, fluency, and the brain* (pp. 333–355). York Press.

Verhoeven, L., & van Leeuwe, J. (2012). The simple view of second language reading throughout the primary grades. *Reading and Writing, 25*(8), 1805–1818. https://doi.org/10.1007/s11145-011-9346-3

Wanzek, J., Roberts, G., Linan-Thompson, S., Vaughn, S., Woodruff, A., & Murray, C. S. (2010). Differences in the relationship of oral reading fluency and high-stakes measures of reading comprehension. *Assessment for Effective Intervention, 35*(2), 67–77. https://doi.org/10.1177/1534508409339917

Wise, J. C., Sevcik, R. A., Morris, R. D., Lovett, M. W., Wolf, M., & Schwanenflugel, P. (2010). The relationship between difference measures of oral reading fluency and reading comprehensions in second-grade students who evidence different oral reading fluency difficulties. *Language, Speech, and Hearing Services in Schools*, *41*(3), 340–348. https://doi.org/10.1044/0161-1461(2009/08-0093)

Wood, D. (2006). Modeling the relationship between oral reading fluency and performance on a statewide reading test. *Educational Assessment*, *11*(2), 85–104. https://doi.org/10.1207/s15326977ea1102_1

Wu, X., & Anderson, R. C. (2007). Reading strategies revealed in Chinese children's oral reading. *Literacy Teaching and Learning*, *12*(1), 47–72.

Wu, Q., Hong, W., & Deng, S. (2017). Application of Chinese character identification in placement tests for CSL learners: An empirical study of constructing simple Chinese proficiency indicators. *Chinese Teaching in the World*, *3*, 395–411.

Xue, J., Shu, H., Li, H., Li, W., & Tian, X. (2013). The stability of literacy-related cognitive contributions to Chinese character naming and reading fluency. *Journal of Psycholinguistic Research*, *42*(5), 433–450. https://doi.org/10.1007/s10936-012-9228-0

Yao, Y. (2011). Interword spacing effects on reading Mandarin Chinese as a second language. *Writing Systems Research*, *3*(1), 23–40. https://doi.org/10.1093/wsr/wsr009

Zhang, J., McBride-Chang, C., Wong, A. M., Tardif, T., Shu, H., & Zhang, Y. (2014). Longitudinal correlates of reading comprehension difficulties in Chinese children. *Reading and Writing*, *27*(3), 481–501. https://doi.org/10.1007/s11145-013-9453-4

Zhou, W., Ye, W., & Yan, M. (2020). Alternating-color words facilitate reading and eye movements among second-language learners of Chinese. *Applied Psycholinguistics*, *41*(3), 685–699. https://doi.org/10.1017/S0142716420000211

Ziegler, J. C., Bertrand, D., Toth, D., Csepe, V., Reis, A., Faisca, L., Saine, N., Lyytinen, H., Vaessen, A., & Blomert, L. (2010). Orthographic depth and its impact on universal predictors of reading: A cross-language investigation. *Psychological Science*, *21*(4), 551–559. https://doi.org/10.1177/0956797610363406

Appendices

Appendix A: Word reading

	1	2	3	4	5
1	我	材料	本来	采访	孩子
2	先	星期	来信	举办	设计
3	礼物	比较	水平	之外	专家
4	电话	为主	大	合上	眼泪
5	课	老师	保护	传统	于是

Appendix B: Word Segmentation

1 放假以后我先休息几天，然后去一个离家很近的地方打工，挣点儿钱，上大学以前去中国旅游两个星期。

2 老师说你住在中国的首都，我住在美国的首都。我喜欢吃中国饭，写汉字，画中国画，尤其爱练中国武术。
3 大球里边的那些比较轻的东西往上升，变成了天；比较重的东西往下降，变成了地，天和地就这样分开了。

4 我下午来找你，可是你不在家。我想邀请你明天跟我去参加我校举办的一年一度的郊游野餐。

5 专家们还认为，现在中国学生常常听的是一些格调不高的流行音乐，这会让他们的思想受到不良影响。

6 你明天早上来了以后，可以先去看一下我们的图书馆（在学校大门的左边），然后还可以去看看图书馆后面的科学实验楼。

Appendix C: Text reading

我下午来找你，可是你不在家。我想邀请你明天跟我去参加我校举办的一年一度的郊游野餐，每位参加者要带一份十元以内的小礼物。餐后有抽奖活动，我想一定很好玩。你回来后尽快给我打个电话。

你好！好久没跟你联系了。最近学习忙吗？我今年学习非常紧张，另外准备申请大学的材料也花了很多时间，昨天刚刚把申请信寄出去。好在我们就要放假了。放假以后我先休息几天，然后去一个离家很近的地方打工，挣点儿钱，上大学以前去中国旅游两个星期。你上次来信说你申请了五所大学，都是哪些大学？预祝成功。

对不起。本来想明天上午陪你参观我们的学校，可是我下午的课临时改到上午了。你明天早上来了以后，可以先去看一下我们的图书馆（在学校大门的左边），然后还可以去看看图书馆后面的科学实验楼。我们的教学楼在学校大门的右边，我十二点下了课后在教学楼门口等你。下午我陪你去参观学校的体育馆和其他地方。

我的中文老师告诉我们，中国学生的英文水平比美国学生的中文水平高多了。请你来信一定要给我介绍你学习外语的“成功秘密”。明年春假，我要跟老师去中国旅行，我真想早日看到长城，北京的故宫，河南的少林寺，四川的熊猫保护基地，特别是西安的兵马俑。

据了解，通过和欧美大学生的比较，教育部的专家发现中国学生学习的课程大部分都是和专业直接有关的，对自己专业之外的东西了解极少，特别是在艺术方面，常常缺乏最起码的知识。欧美大学常常要求学生选修和专业无关的课程。例如一个学化学的学生必须学习五门人文科学课程。这样培养出来的学生，就不是简单的技术人员，而是全面的人才。

他担心天和地会再合上，就站在天和地的中间，用头顶着天，用脚踩着地。天和地之间的距离越来越大，这个孩子也越长越高，越长越壮，变成了一个巨人。这个巨人就这样站着，好像一根大柱子。过了一万八千年，巨人觉得实在太累了，再也坚持不住了，于是他就倒了下去。只听到“轰”的一声，他的头发变成了树林，肩膀变成了高山，肚子变成了平原，汗水变成了大海，血液变成了湖泊，口水变成了河流，眼泪变成了雨水，嘴里呼出的气变成了风和云。这样才有了这个美丽的世界。

About the Author

Shuyi Yang is an instructor at Johns Hopkins University. Her research interests include cognitive processes in Chinese L2 reading comprehension, Chinese L2 reading assessment, and Chinese L2 vocabulary acquisition.
E-mail: shuyi.yang@jhu.edu